THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Detailed Molecular and Immune Marker Profiling of Archival Prostate Cancer Samples Reveals an Inverse Association between TMPRSS2:ERG Fusion Status and Immune Cell Infiltration

Citation for published version:
Rao, SR, Alham, NK, Upton, E, Mcintyre, S, Bryant, RJ, Cerundolo, L, Bowes, E, Jones, S, Browne, M, Mills, I, Lamb, A, Tomlinson, I, Wedge, D, Browning, L, Sirinukunwattana, K, Palles, C, Hamdy, FC, Rittscher, J & Verrill, C 2020, 'Detailed Molecular and Immune Marker Profiling of Archival Prostate Cancer Samples Reveals an Inverse Association between TMPRSS2:ERG Fusion Status and Immune Cell Infiltration', *The Journal of Molecular Diagnostics*, vol. 22, no. 5, pp. 652-669. https://doi.org/10.1016/j.jmoldx.2020.02.012

Digital Object Identifier (DOI):
10.1016/j.jmoldx.2020.02.012

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
The Journal of Molecular Diagnostics

Publisher Rights Statement:
This is a pre-copyedited, author-produced version of an article accepted for publication in The Journal of Molecular Diagnostics following peer review. The version of record "Detailed Molecular and Immune Marker Profiling of Archival Prostate Cancer Samples Reveals an Inverse Association between TMPRSS2:ERG Fusion Status and Immune Cell Infiltration" is available online at:https://doi.org/10.1016/j.jmoldx.2020.02.012

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 06. May. 2023

Detailed molecular and immune marker profiling of archival prostate cancer samples reveals an inverse association between *TMPRSS2*:*ERG* fusion status and immune cell infiltration

Running title: Immune profiling in prostate cancer

Srinivasa R. Rao[1], Nasullah K. Alham[2,3], Elysia Upton[1], Stacey McIntyre[1], Richard J. Bryant[1], Lucia Cerundolo[1], Emma Bowes[1,3], Stephanie Jones[1], Molly Browne[1,3], Ian Mills[1], Alastair Lamb[1], Ian Tomlinson[4], David Wedge[2,3], Lisa Browning[3,5], Korsuk Sirinukunwattana[2], Claire Palles[4], Freddie C. Hamdy[1], Jens Rittscher[2,$], Clare Verrill[1,3,5,$#]

$ Contributed equally
# Corresponding author

Affiliations
1 Nuffield Department of Surgical Sciences, University of Oxford, Oxford, OX3 9DU
2 Big Data Institute, University of Oxford, Old Road Campus, Oxford, OX3 7LF
3 Oxford NIHR Biomedical Research Centre, Oxford, OX4 2PG
4 Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT
5 Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU

Manuscript details
number of text pages: 30
tables: 2
figures: 8

Corresponding author:
Clare Verrill
Phone: +44 (0)1865 220150
Email: clare.verrill@nds.ox.ac.uk

Funding Statement
CRUK programme grant: C1380/A18444
CRUK Dev Fund is: CRUKDF-0715-CV
The research was also funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health (CV/JR).

Disclosure/Conflict of Interest:
The authors confirm that there is no conflict of interest to disclose.

Abstract:

Prostate cancer is a significant global health issue and limitations to current patient management pathways often result in over- or under-treatment. New ways to stratify patients are urgently needed. We conducted a feasibility study of such novel assessments looking for associations between genomic changes and lymphocyte infiltration. An innovative workflow utilising an in-house targeted sequencing panel, immune cell profiling using an image analysis pipeline, RNA-Seq, and exome sequencing in select cases was tested.

Gene fusions were profiled by RNA-seq in 27/27 cases and a significantly higher TIL count was noted in tumours without a *TMPRSS2*:*ERG* fusion compared to those with the fusion (p=0.01). Although this finding was not replicated in a larger validation set (n=436) of The Cancer Genome Atlas images, there was a trend in the same direction. Differential expression analysis of TIL-High and TIL-Low tumours revealed the enrichment of both innate and adaptive immune response pathways. Mutations in mismatch repair genes (*MLH1* and *MSH6* mutations in 1/27 cases) were identified.

We describe a potential immune escape mechanism in *TMPRSS2*:*ERG* fusion positive tumours. Detailed profiling, as shown here, can provide novel insights into tumour biology. Likely differences with findings with other cohorts is related to methods used to define region of interest, but this warrants further study in a larger cohort.

Keywords: inflammation; next generation sequencing; digital image analysis; immunohistochemistry.

Introduction:

Prostate cancer is a major global health issue, and the second most common cause of cancer death in males in the developing world[1]. Better methods for patient stratification are urgently needed.

Our understanding of the molecular pathology of prostate cancer is evolving fast with advances in sequencing methods and bioinformatics. The Cancer Genome Atlas (TCGA) performed detailed molecular analysis on 333 primary prostate carcinomas and 74% of tumours fell into one of 7 subtypes defined by specific gene fusions (*ERG*, *ETV1/4*, *FLI1*) or mutations (*SPOP*, *FOXA1*, *IDH1*)[2]. Recent publications by the International Cancer Genome Consortium have identified new cancer genes, routes of progression and drug targets, sometimes affected by mutations in non-coding regions of genes including *NEAT1* and *FOXA1*[3].

There is now greater understanding of the links between morphology and molecular alterations (specific morpho-molecular correlations), molecular aberrations and inflammation and the therapeutic implications of these associations. Immunotherapy is a treatment option for late stage prostate cancer and in order to optimise response to immunotherapy, it is often necessary to target oncogenic driver pathways in combination with immunotherapy[4]. Increasing evidence indicates the clinical utility of inflammatory infiltrates in determining cancer prognosis, with a widely accepted role for the immune system in controlling cancer growth and progression. In a number of different cancers including prostate, colorectal, melanoma and bladder, strong T cell infiltration is associated with a favourable clinical outcome[5].

The literature detailing prostate cancer, progression and inflammation is often conflicting. For example, a critical review showed radical prostatectomy cases with higher rates of biochemical progression to have higher levels of systemic inflammatory markers. In another study, higher grade inflammation was statistically associated with risk of extraprostatic extension, positive margins and seminal vesicle invasion[6], which is largely contradictory to the literature in other tumour types. Issues around accuracy of quantification of inflammatory cells, which in many of these studies will have been conducted by pathologist assessment, can be addressed with quantitative image analysis.

Prostate cancer shows a high degree of heterogeneity at histological and genetic levels, which poses a significant treatment challenge. As mutations and activation of oncogenic driver pathways accumulate with tumour progression, this may promote increases in immunosuppressive cells and exhaustion of immune effector cells in the tumour microenvironment[4].

A recent study showed 2 of 25 sequenced prostate cancers to be MMR deficient (1 related to Lynch syndrome, 1 sporadic) and showed high inflammatory infiltrates and loss of the 2 relevant MMR proteins on IHC (*MSH2* and *MSH6*)[7]. A study describing prostate cancer in Lynch syndrome showed tumours to be generally high grade (Gleason Scores 8-10) with mutations in *MSH2*, *MLH1* and *MSH6* and loss of the respective protein on IHC in 69% of tumours[8]. MSI and dMMR have been reported in prostate cancers to range from approximately 1% in primary up to 12% of metastatic cancers[9].

Other associations that have been found between genomics and inflammation include; the loss of *PTEN* and immunosuppression in a dose dependent manner and that the loss or mutation of p53 may enhance the immunosuppressive microenvironment[4]. At fusion level, it

has been postulated that *TMPRSS2*:*ERG* fusions generate chimeric amino acid sequences that are targetable by T cells[10]. Prediction algorithms have been used to identify potentially antigenic epitopes from *TMPRSS2*:*ERG* fusions[11].

Here, we aimed to study the relationship between immune cell infiltration and mutations in prostate cancer by using an integrative approach combining image analysis, DNA and RNA sequencing.

Methods:

Workflow

This study involved digital image analysis or next generation sequencing of archival prostate cancer samples as depicted in Figure 1.

Ethics

The study was undertaken under the Oxford Radcliffe Biobank Research Tissue Bank REC approval (reference 09/H0606 5+5) and informed consent was obtained in all cases.

Cohort

27 unselected and sequential radical prostatectomy cases stored as FFPE blocks in the diagnostic archives of Oxford University Hospitals NHS Foundation Trust. The cases dated from 2014-2016 and sequencing was conducted in 2017 and 2018, thus the range of age of blocks (samples) was between 1 and 4 years. 18 cases were Gleason Grade Group 2 (Gleason Score 7 (3+4)), 7 cases were Grade Group 3 (Gleason Score 7(4+3)) and one of these cases had tertiary pattern 5 and 2 cases were Grade Group 5 (Gleason Score 9 (4+5)). 11 cases were

prostate confined tumour (stage pT2), 13 showed extraprostatic extension (pT3a) and 2 showed seminal vesicle invasion (pT3b).

Staining

Tissue sections were subject to routine H&E staining, using a Tissue-Tek Prisma Autostainer (Sakura, Flemingweg, Netherlands). Immunohistochemistry was performed using a Bond RX automatic stainer (Leica Biosystems, Wetzlar, Germany) using the IHC Protocol F program, following incubation of sections at 60°C for 10 minutes. This machine performed all steps including de-paraffinisation, antigen retrieval and staining, counterstaining and washing using Leica Biosystems reagents: Leica Bond Dewax Solution (AR9222), Wash Buffer (AR9590), epitope retrieval solution 1 (AR9961), epitope retrieval solution 2 (AR9640), enzyme 1 from the enzyme pre-treatment kit (AR9551), and Bond Polymer Detection System (DS9800). Epitope retrieval conditions were dependent on the antibody manufacturer's recommendations (Table 1). Antibodies requiring dilution prior to use (anti-FoxP3, anti-PMS2 and anti-MSH6) were diluted to 1:100 using Bond Primary Antibody Diluent (Leica Biosystems, AR9352). For all antibodies, tonsil tissue sections were used as a positive control. Following protocol completion, sections were dehydrated with serial 1 minute washes in 70%, 90% and twice in 100% ethanol (Sigma Aldrich, St Louis, Missouri, USA - 652261). This was followed by 2x 5 minute histo-clear II (National Diagnostics, USA HS-202) washes and mounting using omnimount (National Diagnostics, HS-110).

Immunohistochemistry was performed on FFPE tissue with antibodies to CD3, CD4, CD8, CD20, Granzyme B, FoxP3, MCK, CK5, MSH2, MSH6, MLH1 and PMS2 on whole mount sections and PTEN, AR and ERG on a TMA. Full details of the staining protocols and primary antibodies are shown in Table 1. The Bond Polymer Refine Detection kit, which utilises DAB as the

chromogen and haematoxylin as the counterstain, was used as the detection system on a Leica Bond Immunostainer (Leica Biosystems). Tonsil tissue was used as a positive control for all markers.

Immune Cell Quantification by Digital Image Analysis (DIA)

Stained sections were scanned at 20x using a NanoZoomer 2.0 digital pathology slide scanner (Hamamatsu Photonics, Hamamatsu-City, Japan). Manual visual analysis was performed using the Hamamatsu NDP viewer (Hamamatsu Photonics, U12388-01). Digital Image Analysis (DIA) was performed using Visiopharm's Integrator System platform (Hoersholm, Denmark), version number 2018.4.3.4480 and image analysis protocols were implemented as Analysis Protocol Packages (APP).

Several APPs were designed to quantify and calculate TIL density on slides stained with CD3, CD4, CD8, CD20, FOXP3, GRANB and H&E. The Image analysis process consists of the following steps:

1. Image Alignment/Registration: The Tissuealign module was used to align 3 digitised serial sections; two slides stained with tumor markers CK5 and PanCK and another slide stained with a cancer biomarker e.g. CD3, CD4, CD20, etc. The alignment was performed both on a large scale, and on a finer detailed level, to get the best possible match of the 3 tissue sections. The alignments were verified visually.

2. Detecting Region of Interest (ROI): For invasive tumour detection, a region of interest (tumour) detection APP was created. The APP identified panCytokeratin (MCK)+ luminal epithelial cells and CK5+ basal cells. Basal cells are absent in prostate adenocarcinoma glands, but present in benign glands (or prostatic intraepithelial neoplasia – PIN). Differential staining

was used to define regions of interest (MCK+ CK5+ regions = 'non tumour', MCK+ CK5- regions = 'invasive tumour'). In addition, the tumour region was also additionally divided into invasive margin and centre of tumour using an APP built in house using the Erode post-processing function of the VIS platform that decreases the tumour region by 120 pixels. The first auxiliary APP was run on the slide using threshold classification that identifies the tissue regions. The second auxiliary APP was run on the PanCK slide using threshold classification that identifies the tumour regions, invasive margin and centre of tumour. The third auxiliary APP was run on the CK5 slide using threshold classification that identifies the benign regions. The ROIs were then superimposed on the aligned cancer biomarker slide to outline various regions for subsequent analysis limited to the inside of the specific regions. The ROI detection APPs operates at a low magnification which enables outlining ROIs in a few seconds. We used 50 patches (0.8mm X 0.8mm) that were generated from 5 representative images to select the optimal threshold. Supplementary figure 1 shows the output of the ROI detection pipeline.

3. Immune Cell Quantification: HDAB-DAB colour deconvolution band was used to detect positively stained cells on CD3, CD4, CD8, CD20, FOXP3 and GRANB slides. H&E-Haematoxylin colour deconvolution band was used to detect positively stained lymphocytes on H&E slides. To enhance the stained cells, while suppressing the background variation, several pre-processing steps were included. The colour deconvolution bands were inputted into a threshold classifier. Thresholding classification method defines a threshold for a given feature, and assigns one class to all pixels with a feature value above or equal to that value, and another class for the rest.

The classification rule was defined as:

$$Class\big(Feature(X,Y)\big) = \begin{cases} A, Feature(x,y) \geq T \\ B, \quad Otherwise \end{cases}$$

where T is the user-selected threshold (cut-off value), A and B are the labels/classes to which the pixel is assigned.

As post-processing steps; a method for cell separation which is based on shape and size was used, cell areas that were too small were removed and finally unbiased counting frames were applied to avoid cells that were intersecting with neighbouring tile boundaries and avoid them being counted twice (or more). The APPs operate at magnification 10X which enables analysing a whole slide image in 5-7 minutes. All immune marker results discussed are density as calculated by number of cells detected / area (mm^2).

Quantitative Output Variables and Calculations

The output variables obtained from the APPs are shown in Table 1. Validation of APPs for the quantification of immune infiltrates was achieved by performing a comparison between the APP and manual cell counting of equivalent images by a pathologist showed good concordance (Supplementary table 1).

TIL density was also calculated in The Cancer Genome Atlas data set (downloaded from portal.gdc.cancer.gov/repository, accessed on 25/08/2019) using the Visiopharm H&E TIL APP. As IHC sections are not part of the dataset, ROI was defined by manual pathologist annotation on the AIDA platform (https://imageannotation.nds.ox.ac.uk:8443/AIDA/ accessed on 24/08/2019)

Scoring and Morphological Assessments

The majority of scoring was done with image analysis as described above. For some of the stains (Mib-1, PTEN, PDL-1, AR, ERG), manual pathologist scoring was undertaken as the staining pattern was complex and needed some degree of pathologist interpretation. For PTEN and AR, H scoring was undertaken. This is a commonly used scoring system, using the following formula [1 × (% cells 1+) + 2 × (% cells 2+) + 3 × (% cells 3+)] with a final score of 0-300. For PTEN, in addition, guidance available in Ferraldeschi et al. was used, in which an H-Score of <10 was considered negative and also cases were separately assessed for clonally negative areas with any negative areas being regarded overall as negative[12]. No standard definition exists for PTEN positivity or loss by IHC and thus we followed this previously described system. In that study, they showed that all cases with a homozygous loss of PTEN had loss on IHC, heterozygous loss cases had H Scores of (median) 0-80 (low/absent) and in heterozygous loss cases, there was commonly loss of IHC expression. PDL-1 was scored according to guidance previously published[13] where staining is scored as negative (0), weak (1), moderate (2) or strong (3) for specific membrane and cytoplasmic staining of epithelial tumor cells. ERG was assessed by a previously described method[14] of intensity only scoring. Cases are scored from 0-3 by visual assessment and then cases that are scored as 1-3 are considered positive (ERG-high) with only cases scored as 0 being considered negative (ERG-low).

Tumour budding was defined as previously described in colorectal carcinoma studies as single cells or clusters of up to 4 or 5 cells at the invasive tumour front[15]. Assessment was made using standard H&E slides, but if assessment was difficult, this was further facilitated by using panCK stains. The border of the tumour was also categorised as being a pushing (or expanding) or infiltrating border. A pushing tumour border was described as one with margins

which were reasonably well circumscribed, whereas an infiltrative tumour border was described as one with dissection of normal tissue with loss of a clear boundary between tumour and host tissues[15] (Supplementary table 2).

The presence or absence of an intraductal adenocarcinoma and if present, percentage volume was assessed by a pathologist (CV). If it was unclear from the H&E section, the CK5 stain was referred to for clarification.

DNA Sequencing

Targeted deep sequencing

The region of tissue to be used for DNA extraction was determined by pathologist assessment of H&E-stained slides and tissue macrodissection was performed. DNA was extracted using the High Pure FFPET DNA Isolation kit (Roche, Basel, Switzerland, #06650767001). DNA was quantitated using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, Massachussetts, USA #Q32854) and concentration ranged between 2.3 - 73.5 ng/ul. 50 ng of template DNA was used for all samples except one (in which the concentration was lower than 2.5 ng/ul). DNA SmMIP sequencing was performed as previously described[16]. Briefly, single-molecule Molecular Inversion Probes (smMIPs) were designed using MIPgen[17] and were synthesised by Integrated DNA Technologies, Coralville, Iowa, USA. Initially, a capture reaction was performed whereby smMIPs anneal to target sequences (50 ng of template DNA), and gap filling occurs using the intervening target DNA as a template. Exonuclease treatment is then used to remove all linear DNA, leaving circularised smMIPs. PCR using internal primer sites amplification of the probes, and sequencing is followed by consensus generation for the

genes of interest. Sequencing was performed at a depth of 2000X, to give an effective depth >150X after de-duplication. Depth of coverage per probe is shown in Supplementary figure 2.

The MIP sequencing panel consisted of 3189 MIPs, covering coding exons or hotspots in 23 genes (*AR*, *ARID1A*, *CDK12*, *CHD1*, *CTNNB1*, *FOXA1*, *IDH1*, *KDM6A*, *MED12*, *MLH1*, *MSH2*, *MSH6*, *NCOR1*, *NCOR2*, *NKX3-1*, *PIK3CA*, *PIK3CB*, *PIK3R1*, *PTEN*, *RB1*, *SPOP*, *TP53*, *ZFHX3*) and spanning ~78 kbp on both strands with an approximate coverage of ~95%. Genes were selected based on a literature review, which included commonly reported prostate cancer driver genes[18] *PIK3CA* and *TP53* are potential targets of investigational drugs, *CTNNB1*, *IDH1*, *NCOR1*, *NCOR2* and *PIK3R1* are potential targets being investigated chemically[3].

Whole exome sequencing

Whole exome sequencing was performed in selected samples using the TruSeq DNA Exome kit (Illumina, San Diego, California, USA, #20020614) with the following modifications: A total of 100-300ng DNA was used for the library prep, depending on sample quality, pre-enrichment amplification was done for 12 cycles, and 500ng amplified DNA was taken forward for enrichment. Samples were multiplexed and run on a NextSeq 500 sequencer (Illumina) as paired-end at 75bp read length, at a minimum average read depth of 70X for tumour samples and 35X for matched normal (after filtering PCR duplicates).

Variant calling

Targeted sequencing

For targeted deep sequencing, samples were first deduplicated based on the unique molecular index sequence. Somatic variants (SNVs and indels) were called using Lofreq*[19]. C>T and G>A mutations occurring only on one strand of the template, with a wild-type allele

on the complementary strand were considered to be formalin-fixation or PCR artefacts, and hence excluded from the list of true mutations.

Exome sequencing

For whole exome sequencing, somatic variant calling was performed according to the GATK (v4.0.6.0)[20] Best Practices workflow and further filtered using custom filters (tlod > 10, SOR < 2). Briefly, this involved filtering of PCR duplicates, recalibration of base quality, and somatic mutation calling with matched normals using Mutect2[21].

In both the above analyses, variants were annotated using the Variant Effects Predictor (v91.3) (VEP)[22].

RNA Sequencing and analysis

RNA was extracted from pathologist-marked unstained formalin-fixed paraffin embedded radical prostatectomy sections as for DNA, using the High Pure FFPET RNA Isolation kit (Roche, Basel, Switzerland #06650775001). RNA was quantitated using the Qubit RNA HS Assay Kit (ThermoFisher Scientific, #Q32855). The quality of the RNA was checked using a High Sensitivity RNA Screentape (Agilent, Santa Clara, California, USA, #5067- 5579) on a Tapestation 2200 (Agilent), with RNA Integrity Numbers ranging from 1 to 4.

The RNA was depleted of abundant ribosomal transcripts using the NEBNext rRNA Depletion Kit (NEB, #E6310L) and libraries were prepared using the NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB, #E7770S) according to the manufacturer's recommendations. The resultant libraries were multiplexed and run on a NextSeq 500 sequencer (Illumina) as paired-end at 75bp read length, at an average of 20 million reads per sample.

Reads were demultiplexed, aligned to the reference genome (hg38) using STAR aligner (v2.5.0)[23] and fusions were called using STAR-Fusion (v1.4.0)[24]. The called fusions were graphically visualised using the Chimeraviz R package (v1.5.6)[25].

For TIL differential expression analysis, median TIL density in the centre of the tumour (H & E) was used as a cut-off to stratify patient samples into TIL-High (n = 11) or TIL-Low (n = 12). Samples for which TIL counts were not available (n = 4) were excluded. Read counts per transcript were generated using Salmon (v0.11.3)[26], summarised at gene level using tximport (v1.6.0)[27] and differential expression analysis was performed using DESeq2 v(1.18.1)[28]. For differential expression analysis based on TMPRSS2:ERG fusion status, samples were stratified into fusion-positive (n = 10) or fusion-negative (n = 17) based on STAR-Fusion calls. The rest of the differential expression analysis was performed as described above.

For gene set enrichment analysis, DESeq2-normalised (based on library size) counts were provided as input to GSEA (v3.0)[29] and tested for enrichment on the KEGG gene sets (from the Molecular signatures database MSigDB[30]) at a nominal p value threshold of 0.05 and FDR threshold of 0.25. 'Signal2Noise' metric was used for ranking genes.

HLA typing

HLA types of all samples were obtained from STAR-aligned RNA-seq data using seq2HLA (v2.3)[31].

Neoantigen prediction

Neoantigens from fusion chimeras were predicted using the INTEGRATE-Neo pipeline (v1.2.1)[11], again using hg38 as the reference genome, followed by pvacfuse (pvactools v1.0.7). Neoantigens from SNVs and short indels were predicted using pvacseq (pvactools v1.0.7)[32].

In both cases, predictions were made for peptides of 8-11 amino acids length for MHC class I and 15 amino acids length for MHC class II. NetMHCpan[33] and NetMHCIIpan[34] were used as the prediction algorithms for MHC class I and class II respectively.

MSI analysis

SmMIP targeted sequencing data from matched tumour and normal samples was analysed with MSIsensor[35].

Statistical Analysis

All statistical analyses were performed in R statistical software (v3.4.4). All tests were performed at a significance level of $\alpha = 0.05$, except for FDR < 0.25 for GSEA (according to software recommendations).

Results:

Automated quantification of immune cell infiltrates

Immune cell infiltration in tumour tissue was quantified on IHC tumour sections using the Visiopharm platform (Figure 2). Areas of invasive tumour were identified using MCK and CK5 staining (see methods for additional details), and then further APPs to quantify immune cells were applied. Additional stains and APPs were used to further subclassify the infiltrating immune cells into helper (CD4+), cytotoxic (CD8+), regulatory (FOXP3+), activated T (Granzyme B+) cells and B (CD20+) cells. In addition, the tumour region was subdivided into invasive margin (IM) and the centre of the tumour (CT) based on proximity to benign tissue and guidance in previously published literature[36, 37]. This analysis revealed that the number of immune cells within the prostate tumours varied across samples, as well as across immune cell subtypes (Figure 3A, B). A moderate to high positive correlation was observed between the numbers of different T cell subtypes, particularly among CD3, CD4 and CD8 counts. A high negative correlation was seen only between regulatory T cells (FOXP3+) and B (CD20+) cells in the invasive margin (Figure 4A). In addition, a high positive correlation was observed between TIL counts in IM and CT regions (Pearson's $R = 0.86$, $p < 0.001$) (Figure 4B) whereas a lower correlation was observed between TIL counts in benign and CT regions (Pearson's $R = 0.41$, $p = 0.05$) (Figure 4C).

In 22 cases the ROIs were successfully identified by DIA however in 5 cases manual drawing was needed due to a complex tumour shape or poor staining on the MCK and CK5 slides. The pipeline was also tested on a TMA as well as the whole slide images used for the main study and was able to run successfully on most cores (Supplementary table 3).

Identification of gene fusions and correlation with immune cell infiltrates and IHC

Whole transcriptome RNA sequencing was successfully performed on RNA extracted from the 27 FFPE radical prostatectomy samples, to identify gene fusions. In the context of prostate cancer, ETS family gene fusions were of particular interest. Fusions involving any ETS family gene as the downstream partner were found in 12/27 samples (44%), of which 10/27 were involving *TMPRSS*:*ERG*, 1/27 was *TMPRSS2*:*ETV4* and 1/27 was *SLC45A3*:*ERG*. In addition, other previously reported gene fusions such as *SLC45A2*:*AMACR* and *EIF4E3*:*FOXP1* were also identified (Table 2). Transcripts and exons involved in *TMPRSS2*:*ERG* fusions are shown in Supplementary figure 3.

On IHC scoring, 26 samples (1 excluded in IHC scoring due to no tumour being present), 11 samples were ERG-high (score > 0) and 25 were ERG-low. 9 samples were fusion positive by RNA Seq and 17 were fusion-negative. 2/9 fusion positive samples were ERG-low. 4/17 fusion-negative samples were ERG-high (Figure 5A). Other fusions were identified with TMPRSS2, EU_H had a TMPRSS2:ETV4 fusion. Gene fusions lead to the formation of neopeptides that may act as tumour neoantigens. RNA Seq data (above) were used to determine the HLA type (MHC class I and II) of the 27 patients. This information combined with the fusion status was used to predict the antigenicity of the tumour neopeptides. Table 1 shows the list of fusions and the samples for which neoantigens are predicted (binding affinity < 500 nM) (Supplementary table 4). The most striking result was that total immune cell counts (identified by H & E) in the centre of the tumour were significantly higher ($p = 0.01$, Wilcoxon test) in the *TMPRSS2*:*ERG* fusion negative compared to fusion positive samples (Figure 5B).

We attempted to validate this finding in the TCGA dataset (n=436) using our Visiopharm H&E APP, with tumour regions manually marked by a pathologist from H&E images (IHC

stains do not form part of this dataset). There was a very small difference between *TMPRSS2*:*ERG* fusion positive and negative cases, but it was not significant (p-value = 0.5474) (Supplementary figure 4A). However, importantly the trend was in the same direction as our finding of *TMPRSS2*:*ERG* fusion and low TILS. Results were similar when looking at immune cell density in the centre of the tumour, invasive margin, or all of the tumour. 13 samples had been excluded from this analysis due to the low quality of the images. Further, a tissue microarray from the CamCap cohort[38] (n = 242) was analysed for lymphocyte density using our Visiopharm H&E app. There was no significant association between lymphocyte density and TMPRSS2:ERG fusion status in this TMA (p = 0.77, Welch's two sample t test) (Supplementary figure 4B).

In order to determine the TIL signature, the samples were stratified into TIL-High (n = 11) and TIL-Low (n = 12) PCa, using the median TIL density in the centre of the tumour (H & E) as the cut-off. Samples with tumour area too small to demarcate a non-zero centre were excluded (n = 4). Differential expression analysis between these two groups of samples identified 188 genes that are significantly (adjusted p value < 0.05) up- or down-regulated in TIL PCa (Figure 5C), of which 84 genes are expressed at least 2-fold higher levels in TIL-High compared to TIL-Low samples. Salient among these genes are several immunoglobulin proteins (*IGLV2-14*, *IGHJ4*, *IGHG2*, *IGHJ3*, *IGHV3-48*), granzymes (*GZMK*, *GZMA*) and cytokines/cytokine receptors (*CXCL10*, *CXCL9*, *CXCR4*). Gene set enrichment analysis of the RNA-seq data revealed that 4 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways highly relevant to immune responses are significantly enriched (p < 0.05, FDR < 0.25) in TIL-High samples: these are the T cell receptor pathway, the B cell receptor pathway, toll-like receptor pathway and genes involved in antigen processing and presentation (Figure 5D). *SPINK1* (serine protease inhibitor

Kazal-type 1) gene is the most significant differentially expressed gene (log2 fold change = 4.43, adjusted p value < 0.001), and also the gene that is most highly expressed in TIL-High samples after SOX9 (Supplementary table 5). ERG gene expression is higher in TIL-Low samples and lower in TIL-High samples (log2 fold change = -2.2, adjusted p value = 0.06), but this difference falls short of the statistical significance threshold of $\alpha = 0.05$. Conversely, differential expression analysis of *TMPRSS2*:*ERG* fusion-positive (n = 10) vs. fusion-negative (n = 17) samples identified *SPINK1*, among other genes, to be highly expressed in fusion-negative samples (Supplementary table 6).

Targeted deep sequencing to identify SNVs and short indels

In order to further identify neoantigens produced by SNVs and short indels, we chose to perform targeted sequencing of 23 prostate cancer driver genes using a single-molecule tagged molecular inversion probe panel that was built in-house (6A). This targeted sequencing was successful in 18/27 matched tumour:normal samples. A schematic of the panel design is show in Figure 6A. Missense single nucleotide variants were identified in 6 genes, and putative copy loss was observed in 17 genes in total (Figure 6B). No mutations were observed in the targeted genes in 6 patients. Identifying copy number alterations poses a significant problem in targeted sequencing, especially in a focused panel such as this. Hence, copy loss of the most frequently deleted genes *PTEN*, *CHD1* and *RB1* was validated using Taqman real-time PCR (Figure 6C). By sequencing and qPCR methods, *PTEN* was lost in 6/18 cases by sequencing loss. However, by IHC, a total of 11/27 cases were found to have a loss of *PTEN*. Only 2/18 cases were found to have loss of *PTEN* both by sequencing as well as IHC.

No neoantigens were predicted from the 6 missense SNVs identified by targeted sequencing.

Whole exome sequencing in select cases revealed potentially coordinated loss of PIK3R1 and CHD1

As very few missense SNVs were identified by targeted sequencing, and no neoantigens were predicted from them, it was difficult to draw any conclusions on the relationship between immune infiltration and SNVs. Hence, whole exome sequencing was performed in 5 selected samples (based on high immune cell infiltrates without explanatory fusions or SNVs in driver genes) for a more comprehensive analysis and to identify structural variants, as well as validate the findings from the targeted sequencing panel. The missense mutation in SPOP was confirmed in sample EU_S, although at a different allele frequency (VAF = 0.38) compared to the targeted sequencing (VAF = 0.70). Interestingly, EU_I had the highest number of mutations (Figure 7A) as well as the highest total immune cell counts as measured by individual markers (Figure 7B). EU_I also had the highest number of neoantigens predicted by NetMHCPan and NetMHCIIpan (binding affinity < 500) (Figure 7C). However, no significant correlation was found (Pearson's R = -0.29, p = 0.63) between the number of predicted neoantigens and the total number of TILs across the 5 samples (Figure 7D). Copy number alterations were identified using GATK4 and are shown as log2 plots of the probes normalised to their matched normal (Supplementary figure 5). Interestingly, this revealed that in 3/5 samples, both *PIK3R1* and *CHD1* appear to be lost due to structural variations in the long arm of chromosome 5. *CDH1* (E-cadherin) was lost in 1/5 sample. In addition, *MYC* was amplified in 2/5 samples and *MAP3K1* (Supplementary table 7). No changes were observed in AR copy number.

Mismatch repair genes mutations

Somatic missense mutations were observed at an allele frequency of 0.28 and 0.20 by targeted deep sequencing in one sample, in *MLH1* (I219V involving the DNA mismatch repair domain) and *MSH6* (S144I) respectively (Figure 8A). To further characterise this, MSH2, MSH6, MLH1 and PMS1 were stained in tissue sections using immunohistochemistry. The mutation in MSH6 was confirmed by IHC as a loss of staining in most of the tumour region whereas the *MLH1* mutation was not corroborated by IHC (Figure 8B). No microsatellite instability was detected in the markers tested, Bat25 and Bat26 (Figure 8C). Further, testing for microsatellite instability in the targeted deep sequencing data identified only 1 somatic site out of 27 microsatellite sites for which sequencing data were available (3.7%) indicating that there is no microsatellite instability in this sample in spite of the mutations in MMR genes. Further, there was no strikingly high immune cell infiltration in this sample.

Other findings

There was no relationship between Gleason Grade Group or pathological stage and either mutations/copy number changes or fusions or any of the inflammatory parameters. Specifically, there was no association between CD3, CD8, CD20, FOXP3, Granzyme B and *TMPRSS2*:*ERG* fusion as had been observed with TILs in the CT as detected on H&E by image analysis. Although correlation was only seen using the H&E APP in the CT, the APP had been validated by manual counting by a pathologist (Supplementary table 2) with close concordance (Pearson's R = 0.99, $p<0.0001$).

Supplementary table 3 shows the scoring results of immunohistochemical staining for PDL-1, PTEN, AR and ERG. PDL-1 IHC showed 23/27 cases to have a score of 0, 3 had a score of 1 and

1 case had a score of 2. No cases scored 3. The proliferation index with Mib-1 was generally low, ranging from 0-8%. AR staining showed a range of H-Scores from 0-300.

3 of the cases showed intraductal spread of adenocarcinoma and of 2 of these (EU_15 and EU_D), it was only a minor component (5% tumour volume), whereas in the third case, the intraductal component was approximately 20% of the tumour volume (EU_07). There were no significant relationships between the degree of budding or infiltration at the leading edge of the tumour and either sequencing or the inflammatory markers (supplemental table 4).

None of the cases that showed seminal vesicle invasion showed *TMPRSS2*:*ERG* fusion.

Discussion:

In this study, we show an inverse association between *TMPRSS2*:*ERG* fusion status and TIL counts in 27 PCa tumours, and identified a differential gene expression profile in TIL-High vs. TIL-Low tumours. We achieved this by utilising a novel workflow which involved the integration of digital image analysis techniques to obtain TIL counts, followed by detailed gene profiling using RNA and DNA sequencing. Although the same trend was found when we attempted to validate this finding in the TCGA images, it was not statistically significant. Our hypothesis is that our smaller dataset of 27 cases had a very detailed image analysis pipeline applied where the image analysis was able to define the ROI on a gland-by-gland basis. This is the novel aspect of our pipeline, to use IHC to define a detailed ROI that can unmask associations that cannot be identified with a very broad pathologist annotation.

Other published studies have shown the opposite result from our study with high TILs being associated with TMPRSS2:ERG fusion, however, these had the limitations of using IHC on TMAs to determine ERG status with manual annotation of ROI[39, 40]. In our study, RNA Seq

was used to specifically identify TMPRSS2:ERG fusion. IHC detects only the expression status of ERG which could be because of fusion or transcriptional activation by other means. We could find no other studies using RNA Seq and TILs. We also used whole mount sections for image analysis quantification and not TMAs, which have inherent problems with sampling and automated ROI detection to a gland level rather than pathologist annotation to a regional level.

Our finding of a statistically significant association between low TILs and *TMPRSS2*:*ERG* fusion remains when the analysis is performed with stratification of samples into *TMPRSS2*:*ERG* fusions producing chimeric transcripts in the coding exons vs fusion-negative or non-chimeric fusions. This raises the intriguing possibility of the down-regulation of immune cell infiltration (immune escape) in those samples with *TMPRSS2*:*ERG* fusions by an unknown mechanism. A recent finding in a mouse model of prostate cancer suggested that the *TMPRSS2*:*ERG* fusion promotes recruitment of regulatory T cells to the tumour site[41]. This could be a mechanism for immune escape. It has been recently reported that tumours with mutations that bind poorly to MHC class II molecules are positively selected[42], which may also explain the low immune cell counts in *TMPRSS2*:*ERG* fusion positive tumours. In any case, the mechanism is unlikely to involve neoantigens arising from chimeric proteins, as that would have resulted in a higher immune cell infiltration in *TMPRSS2*:*ERG* positive tumour. An indirect neoantigen-independent mechanism, whereby *ERG* overexpression (driven by the *TMPRSS2* promoter) in the fusion-positive cases results in suppression of immune response by alteration of cytokine production, cannot be ruled out. Indeed, such a mechanism has been proposed to explain immune cell infiltration in copy number-driven cancers[43]. It is acknowledged that a change in the number of cases in the cohort could change the value of median cut off used for

differential expression analysis. It should also be noted that even though TMPRSS2:ERG status was also negatively associated with CD3 lymphocyte density, this did not reach statistical significance.

DIA was used to successfully and accurately distinguish between invasive tumour and non-tumour, and to quantify inflammatory infiltrates within these regions. Typically, prostate cancer invades between and around benign glands, therefore manual annotation of tumour regions will often include benign glands. Utilising DIA, increases the accuracy, through annotation on an almost gland-by-gland basis. Furthermore, use of DIA in this manner, removes pathologist subjectivity with the inherent problems of inter- and intra-observer variability[44, 45]. One limitation of using CK5 and panCK stains for region of interest detection is that both Prostatic Intraepithelial Neoplasia (PIN) and Intraductal Spread of Adenocarcinoma retain basal cells like benign glands, but in our study, only 3/27 cases showed intraductal spread of adenocarcinoma and usually to a limited degree and therefore unlikely to have a major impact on the DIA results.

This study utilised serial sections with co-registration rather than a multiplex staining technique such as the IHC based method published by Linch et al.[7] or a more widely used immunofluorescence method. The number of stains that we wished to perform was in excess of the number that could be performed with conventional or published IHC or IF methods. In addition, we focused here on TILs, but if the amount of tissue were not a limitation, we would also have sought to quantify M1 and M2 macrophages and tumour infiltrating neutrophils, which have also been found to be important in addition to TILs in tumour progression[6].

Saltz et al. showed that local clustering of TILs may be a more distinctive feature than overall TIL infiltrate in some tumour types and this was also the case with prostate cancer[46]. In this

study, they also demonstrated a brisk band like infiltrate with immune responses forming band like boundaries bordering the tumour at its periphery. By contrast, although we examined this region in the form of the invasive margin, we did not find this same relationship, with inflammatory cells being generally constant across the CT and IM.

It is previously published that chronic inflammation in benign tissue surrounding PCa is common and positively associated with high grade disease, hence we looked at the background benign tissue and found a moderate positive correlation between the degree of background inflammation and the inflammation in the tumour (total TILs), suggesting that the entire prostate is more inflamed when the tumour is more inflamed[4]. The mechanism for this is unclear, and possibly either reflective of a high pre-existing level in the background prostate or an elicited response to the tumour. Interestingly, it was observed that when tumours were heavily inflamed, the inflammation was often centred around benign glands rather than tumour glands (Supplementary figure 6). In colorectal adenocarcinoma, a body of work supports the concept of inflammation at the invasive margin being related to budding and EMT[15], but we did not find any such associations.

In tumour progression, FOXP3 expressing T regulatory cells suppress the immune system, promoting tumour growth, but by contrast cytotoxic T cells (being associated with a direct cytotoxic effect on tumour cells) and T helper cells confer antitumour immunity and are associated with better prognosis in patients with cancer[47]. Although the role of B lymphocytes has been studied less than T cell responses, B cells can play a role in the immune system against tumour[6]. Despite these previously described associations, we did not find any significant results relating to the quantification of subpopulations of TILs.

RNA sequencing was successful in all 27 samples. However, targeted DNA sequencing failed in 33% of samples, due to poor quality of DNA. This poor quality of DNA is expected due to the known deleterious effects of formalin on nucleic acids[48]. These were unselected sequential cases from the diagnostic archives and no special measures had been implemented to optimise fixation. The smMIP panel allowed successful deep sequencing of 23 target genes in 18/27 tumour specimens. The number of SNVs and indels in the targeted genes was low, in support of previously published data on prostate cancer, which is generally regarded as a mutation poor cancer[2].

Using the TIL counts (H & E) to stratify the samples facilitated the identification of an RNA signature in TIL-High samples. While high expression of immune response genes was expected in TIL-High samples, it was interesting to note the strong signature of B cell and antigen processing signalling pathways, in addition to the T cell signalling pathway, in TIL-High PCa samples. This suggests that the anti-tumour immune response in prostate cancer is a combination of innate and adaptive immune systems. This evidence underlines the importance of both MHC class I and class II molecules in tumour immunity in prostate cancer cells. Interestingly, *SPINK1* (serine peptidase inhibitor Kazal type 1), a gene known to play an important oncogenic role in ETS-fusion negative prostate cancer[49], was highly expressed in TIL-High samples as well as in *TMPRSS2*:*ERG* fusion-negative samples.

Analysis of CNAs was also performed using the smMIP panel, as previously described[50, 51], identifying potential deletions of *PTEN*, *CHD1* and *PIK3R1*. However, this analysis is simplistic, and does not account for variation in sequencing depth due to factors such as GC bias, which was borne out by low concordance with IHC data (for PTEN) as well as with CNA calls from whole exome sequencing data. In contrast, targeted deep sequencing was highly sensitive in

identifying SNVs in *MLH1* and *MSH6* genes at a low minor allele frequency. There was no association between CNA and Gleason Grade Group. It was previously described that more aggressive primary tumours tend to show more extensive CNA[2]. However, our cohort comprised mainly Gleason Grade Group 2 cases, which may explain the difference.

From the whole exome sequencing data, one sample (EU_I) was found to have the highest number of predicted neoantigens, which also corresponded to a high number of immune cell infiltrates (sum of subtypes). However, across all samples, no overall correlation was observed between TILs (by H & E) and the number of putative neoantigens.

In this study we did not see the association of dMMR and prominent inflammatory infiltrates as previously reported[7]. Targeted DNA sequencing is insufficient to assess total mutational burden, which has been postulated to be associated with mutations in MMR genes and sensitivity to immune checkpoint inhibitors[9]. No evidence was found for microsatellite instability, which is a consequence of mutations in mismatch repair genes.

In this study, the relationship between the IHC and the sequencing results was not always an exact one. Ong et al. used IHC as an upfront stratifier of intermediate risk prostate cancer followed by differential gene expression profiling and their IHC staining of PTEN did not correlate with concordant changes in their transcript expression (PTEN loss by IHC 51%)[52]. For ERG, we had mismatches between TMPRSS2:ERG fusion by RNA Seq and IHC. 2/9 fusion positive samples were ERG-low and potential explanations are 1) different regions used for IHC vs sequencing, 2) subclone of tumour has TMPRSS2:ERG fusion and IHC may not be sensitive enough to detect this, 3) limited tumour available to assess (e.g. EU_O, supplemental figure 7). 4/17 fusion-negative samples were ERG-high which can be explained by 1) ERG overexpression due to fusion with genes other than TMPRSS2 (e.g. EU_J had a

SLC45A3:ERG fusion) 2) other unknown mechanisms of ERG overexpression (EU_H had a TMPRSS2:ETV4 fusion). Our rate of concordance between IHC and sequencing was therefore 78%. ERG status by IHC did not correlate with any of the immune subpopulations or TILs by H&E assessment. Rates of correlation between ERG IHC and fusion in other studies include, Sung et al.[14] (95%) and Gopalan et al.[53] (83%). None, however, has used RNA Seq as the platform for determining the TMPRSS2:ERG fusion. In our cohort, ERG fusion was identified with genes other than TMPRSS2.

In conclusion, this study demonstrates the wealth of information that can be obtained from pathology archive prostate cancer samples and could be used in discovery, clinical trial or real time diagnostic practice. Detailed morpho-molecular associations can be studied by image analysis and sequencing, and in this study we report a previously unpublished association between TMPRSS2:ERG fusion negative prostate cancer cases and higher levels of TILs. While it has been shown that TMPRSS2:ERG fusion status is not a prognostic feature in the clinical outcome of prostate cancer in the limited evidence available[54, 55], our results suggest that it could have important implications in the biology of tumour immune response, and although we could not validate this in a larger cohort the explanation is likely due to needing a detailed image analysis pipeline on a gland by gland level to discover such associations, which we could not utilise in the validation cohort.

Acknowledgments:

The research was supported and funded by the National Institute of Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and Cancer Research UK.

The views expressed are those of the author (s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Siegel R, Ma J, Zou Z, Jemal A: Cancer statistics, 2014. CA: a cancer journal for clinicians 2014, 64:9-29.
2. Cancer Genome Atlas Research N: The Molecular Taxonomy of Primary Prostate Cancer. Cell 2015, 163:1011-1025.
3. Wedge DC, Gundem G, Mitchell T, Woodcock DJ, Martincorena I, Ghori M, Zamora J, Butler A, Whitaker H, Kote-Jarai Z, Alexandrov LB, Van Loo P, Massie CE, Dentro S, Warren AY, Verrill C, Berney DM, Dennis N, Merson S, Hawkins S, Howat W, Lu Y-J, Lambert A, Kay J, Kremeyer B, Karaszi K, Luxton H, Camacho N, Marsden L, Edwards S, Matthews L, Bo V, Leongamornlert D, McLaren S, et al.: Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nature Genetics 2018, 50:682-692.
4. Bryant G, Wang L, Mulholland DJ: Overcoming Oncogenic Mediated Tumor Immunity in Prostate Cancer. International Journal of Molecular Sciences 2017, 18.
5. Fridman WH, Pagès F, Sautès-Fridman C, Galon J: The immune contexture in human tumours: impact on clinical outcome. Nature Reviews Cancer 2012, 12:298-306.
6. Sciarra A, Gentilucci A, Salciccia S, Pierella F, Del Bianco F, Gentile V, Silvestri I, Cattarino S: Prognostic value of inflammation in prostate cancer progression and response to therapeutic: a critical review. Journal of Inflammation (London, England) 2016, 13.
7. Linch M, Goh G, Hiley C, Shanmugabavan Y, McGranahan N, Rowan A, Wong YNS, King H, Furness A, Freeman A, Linares J, Akarca A, Herrero J, Rosenthal R, Harder N, Schmidt G, Wilson GA, Birkbak NJ, Mitter R, Dentro S, Cathcart P, Arya M, Johnston E, Scott R, Hung M, Emberton M, Attard G, Szallasi Z, Punwani S, Quezada SA, Marafioti T, Gerlinger M, Ahmed HU, Swanton C: Intratumoural evolutionary landscape of high-risk prostate cancer: the PROGENY study of genomic and immune parameters. Annals of Oncology: Official Journal of the European Society for Medical Oncology 2017, 28:2472-2480.
8. Dominguez-Valentin M, Joost P, Therkildsen C, Jonsson M, Rambech E, Nilbert M: Frequent mismatch-repair defects link prostate cancer to Lynch syndrome. BMC Urology 2016, 16:15.
9. Hempelmann JA, Lockwood CM, Konnick EQ, Schweizer MT, Antonarakis ES, Lotan TL, Montgomery B, Nelson PS, Klemfuss N, Salipante SJ, Pritchard CC: Microsatellite instability in prostate cancer by PCR or next-generation sequencing. Journal for Immunotherapy of Cancer 2018, 6:29.
10. Kalina JL, Neilson DS, Lin Y-Y, Hamilton PT, Comber AP, Loy EMH, Sahinalp SC, Collins C, Hach F, Lum JJ: Mutational Analysis of Gene Fusions Predicts Novel MHC Class I-Restricted T cell Epitopes & Immune Signatures in a Subset of Prostate Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research 2017, 23:7596-7607.
11. Zhang J, Mardis ER, Maher CA: INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. Bioinformatics 2017, 33:555-557.
12. Ferraldeschi R, Nava Rodrigues D, Riisnaes R, Miranda S, Figueiredo I, Rescigno P, Ravi P, Pezaro C, Omlin A, Lorente D, Zafeiriou Z, Mateo J, Altavilla A, Sideris S, Bianchini D, Grist E, Thway K, Perez Lopez R, Tunariu N, Parker C, Dearnaley D, Reid A, Attard G, de Bono J: PTEN protein loss and clinical outcome from castration-resistant prostate cancer treated with abiraterone acetate. European Urology 2015, 67:795-802.
13. Gevensleben H, Dietrich D, Golletz C, Steiner S, Jung M, Thiesler T, Majores M, Stein J, Uhl B, Müller S, Ellinger J, Stephan C, Jung K, Brossart P, Kristiansen G: The Immune Checkpoint Regulator PD-L1 Is Highly Expressed in Aggressive Primary Prostate Cancer. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 2016, 22:1969-1977.
14. Sung JY, Jeon HG, Jeong BC, Seo SI, Jeon SS, Lee HM, Choi HY, Kang SY, Choi YL, Kwon GY: Correlation of ERG immunohistochemistry with molecular detection of TMPRSS2-ERG gene fusion. J Clin Pathol 2016, 69:586-592.

15. Zlobec I, Lugli A: Invasive front of colorectal cancer: Dynamic interface of pro-/anti-tumor factors. World Journal of Gastroenterology : WJG 2009, 15:5898-5906.
16. Eijkelenboom A, Kamping EJ, Kastner-van Raaij AW, Hendriks-Cornelissen SJ, Neveling K, Kuiper RP, Hoischen A, Nelen MR, Ligtenberg MJL, Tops BBJ: Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. The Journal of molecular diagnostics: JMD 2016, 18:851-863.
17. Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J: MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics (Oxford, England) 2014, 30:2670-2672.
18. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N: IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods 2013, 10:1081-1082.
19. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N: LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research 2012, 40:11189-11201.
20. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 2010, 20:1297-1303.
21. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology 2013, 31:213-219.
22. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F: The Ensembl Variant Effect Predictor. Genome Biology 2016, 17:122.
23. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15-21.
24. Haas B, Dobin A, Stransky N, Li B, Yang X, Tickle T, Bankapur A, Ganote C, Doak T, Pochet N, Sun J, Wu C, Gingeras T, Regev A: STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv 2017:120295.
25. Lågstad S, Zhao S, Hoff AM, Johannessen B, Lingjærde OC, Skotheim RI: chimeraviz: a tool for visualizing chimeric RNA. Bioinformatics (Oxford, England) 2017, 33:2954-2956.
26. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nature methods 2017, 14:417-419.
27. Soneson C, Love MI, Robinson MD: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 2016, 4:1521.
28. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 2014, 15.
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 2005, 102:15545-15550.
30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011, 27:1739-1740.
31. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, Türeci Ö, Diken M, Castle JC, Sahin U: HLA typing from RNA-Seq sequence reads. Genome Medicine 2012, 4:102.
32. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M: pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Medicine 2016, 8:11.

33. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, Buus S: NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. PLOS ONE 2007, 2:e796.
34. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M: NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics 2013, 65:711-724.
35. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L: MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 2014, 30:1015-1016.
36. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, Tosolini M, Camus M, Berger A, Wind P, Zinzindohoué F, Bruneval P, Cugnenc P-H, Trajanoski Z, Fridman W-H, Pagès F: Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science (New York, NY) 2006, 313:1960-1964.
37. Galon J, Pagès F, Marincola FM, Angell HK, Thurin M, Lugli A, Zlobec I, Berger A, Bifulco C, Botti G, Tatangelo F, Britten CM, Kreiter S, Chouchane L, Delrio P, Arndt H, Asslaber M, Maio M, Masucci GV, Mihm M, Vidal-Vanaclocha F, Allison JP, Gnjatic S, Hakansson L, Huber C, Singh-Jasuja H, Ottensmeier C, Zwierzina H, Laghi L, Grizzi F, Ohashi PS, Shaw PA, Clarke BA, Wouters BG, et al.: Cancer classification using the Immunoscore: a worldwide task force. Journal of Translational Medicine 2012, 10:205.
38. Dunning MJ, Vowler SL, Lalonde E, Ross-Adams H, Boutros P, Mills IG, Lynch AG, Lamb AD: Mining Human Prostate Cancer Datasets: The "camcAPP" Shiny App. EBioMedicine 2017, 17:5-6.
39. Kaur HB, Guedes LB, Lu J, Maldonado L, Reitz L, Barber JR, De Marzo AM, Tosoian JJ, Tomlins SA, Schaeffer EM, Joshu CE, Sfanos KS, Lotan TL: Association of tumor-infiltrating T-cell density with molecular subtype, racial ancestry and clinical outcomes in prostate cancer. Mod Pathol 2018, 31:1539-1552.
40. Flammiger A, Bayer F, Cirugeda-Kuhnert A, Huland H, Tennstedt P, Simon R, Minner S, Bokemeyer C, Sauter G, Schlomm T, Trepel M: Intratumoral T but not B lymphocytes are related to clinical outcome in prostate cancer. APMIS 2012, 120:901-908.
41. Shan L, Ji T, Su X, Shao Q, Du T, Zhang S: TMPRSS2-ERG Fusion Promotes Recruitment of Regulatory T cells and Tumor Growth in Prostate Cancer. Am J Med Sci 2018, 356:72-78.
42. Marty R, Thompson WK, Salem RM, Zanetti M, Carter H: Evolutionary Pressure against MHC Class II Binding Cancer Mutations. Cell 2018, 0.
43. McGrail DJ, Federico L, Li Y, Dai H, Lu Y, Mills GB, Yi S, Lin SY, Sahni N: Multi-omics analysis reveals neoantigen-independent immune cell infiltration in copy-number driven cancers. Nat Commun 2018, 9:1317.
44. Smits AJJ, Kummer JA, de Bruin PC, Bol M, van den Tweel JG, Seldenrijk KA, Willems SM, Offerhaus GJA, de Weger RA, van Diest PJ, Vink A: The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc 2014, 27:168-174.
45. Buisseret L, Desmedt C, Garaud S, Fornili M, Wang X, Van den Eyden G, de Wind A, Duquenne S, Boisson A, Naveaux C, Rothé F, Rorive S, Decaestecker C, Larsimont D, Piccart-Gebhart M, Biganzoli E, Sotiriou C, Willard-Gallo K: Reliability of tumor-infiltrating lymphocyte and tertiary lymphoid structure assessment in human breast cancer. Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc 2017, 30:1204-1212.
46. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J, Cancer Genome Atlas Research N, Shmulevich I, Rao AUK, Lazar AJ, Sharma A, Thorsson V: Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Reports 2018, 23:181-193.e187.

47. Qu X, Tang Y, Hua S: Immunological Approaches Towards Cancer and Inflammation: A Cross Talk. Frontiers in Immunology 2018, 9.
48. Robbe P, Popitsch N, Knight SJL, Antoniou P, Becq J, He M, Kanapin A, Samsonova A, Vavoulis DV, Ross MT, Kingsbury Z, Cabes M, Ramos SDC, Page S, Dreau H, Ridout K, Jones LJ, Tuff-Lacey A, Henderson S, Mason J, Buffa FM, Verrill C, Maldonado-Perez D, Roxanis I, Collantes E, Browning L, Dhar S, Damato S, Davies S, Caulfield M, Bentley DR, Taylor JC, Turnbull C, Schuh A: Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. Genet Med 2018.
49. Tomlins SA, Rhodes DR, Yu J, Varambally S, Mehra R, Perner S, Demichelis F, Helgeson BE, Laxman B, Morris DS, Cao Q, Cao X, Andrén O, Fall K, Johnson L, Wei JT, Shah RB, Al-Ahmadie H, Eastham JA, Eggener SE, Fine SW, Hotakainen K, Stenman U-H, Tsodikov A, Gerald WL, Lilja H, Reuter VE, Kantoff PW, Scardino PT, Rubin MA, Bjartell AS, Chinnaiyan AM: The Role of SPINK1 in ETS Rearrangement-Negative Prostate Cancers. Cancer Cell 2008, 13:519-528.
50. Kumar A, Boyle EA, Tokita M, Mikheev AM, Sanger MC, Girard E, Silber JR, Gonzalez-Cuyar LF, Hiatt JB, Adey A, Lee C, Kitzman JO, Born DE, Silbergeld DL, Olson JM, Rostomily RC, Shendure J: Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. Genome Biology 2014, 15.
51. Neveling K, Mensenkamp AR, Derks R, Kwint M, Ouchene H, Steehouwer M, Lier Bv, Bosgoed E, Rikken A, Tychon M, Zafeiropoulou D, Castelein S, Hehir-Kwa J, Thung DT, Hofste T, Lelieveld SH, Bertens SMM, Adan IBJF, Eijkelenboom A, Tops BB, Yntema H, Stokowy T, Knappskog PM, Høberg-Vetti H, Steen VM, Boyle E, Martin B, Ligtenberg MJL, Shendure J, Nelen MR, Hoischen A: BRCA Testing by Single-Molecule Molecular Inversion Probes. Clinical Chemistry 2017, 63:503-512.
52. Ong CW, Maxwell P, Alvi MA, McQuaid S, Waugh D, Mills I, Salto-Tellez M: A gene signature associated with PTEN activation defines good prognosis intermediate risk prostate cancer cases. The Journal of Pathology Clinical Research 2018, 4:103-113.
53. Gopalan A, Leversha MA, Dudas ME, Maschino AC, Chang J, Al-Ahmadie HA, Chen YB, Tickoo SK, Reuter VE, Fine SW: TMPRSS2-ERG rearrangement in dominant anterior prostatic tumours: incidence and correlation with ERG immunohistochemistry. Histopathology 2013, 63:279-286.
54. Gopalan A, Leversha MA, Satagopan JM, Zhou Q, Al-Ahmadie HA, Fine SW, Eastham JA, Scardino PT, Scher HI, Tickoo SK, Reuter VE, Gerald WL: TMPRSS2-ERG gene fusion is not associated with outcome in patients treated by prostatectomy. Cancer Res 2009, 69:1400-1406.
55. Dal Pra A, Lalonde E, Sykes J, Warde F, Ishkanian A, Meng A, Maloff C, Srigley J, Joshua AM, Petrovics G, van der Kwast T, Evans A, Milosevic M, Saad F, Collins C, Squire J, Lam W, Bismar TA, Boutros PC, Bristow RG: TMPRSS2-ERG status is not prognostic following prostate cancer radiotherapy: implications for fusion status and DSB repair. Clin Cancer Res 2013, 19:5202-5209.

Figure legends

Figure 1. Workflow used in the study

Figure 2. Image analysis pipeline to quantify immune cells

Figure 3. Immune marker profiling using the Visiopharm platform

A) Absolute immune cell counts in all samples in the centre of the tumour (CT) and B) in the invasive margin of the tumour (IM).

Figure 4. Correlations between immune cell subtypes in different regions of the tumour

A) Correlogram of immune cell subtypes identified. Blue or red colour scale indicates positive or negative correlation respectively, with only those correlations with a p value < 0.05 (not adjusted for multiple comparisons) being shown. B) Relationship between immune cell counts in IM vs. CT (Pearson's R = 0.86, $p < 0.001$) and C) benign region vs. CT (Pearson's R = 0.41, p = 0.05) shown as scatter plots.

Figure 5. Relationship between TMPRSS2:ERG fusion and immune cell infiltration

A) ERG overexpression scored by immunostaining largely matches TMPRSS2:ERG fusions detected by RNA-seq (p = 0.37, McNemar's Chi square test). B) H & E absolute count in the centre of *TMPRSS2*:*ERG* fusion negative and positive tumours. (p = 0.01, Wilcoxon test). C) Volcano plot showing genes that are significantly differentially expressed between samples with high TIL density (High-TIL PCa) versus those with low TIL density (Low-TIL PCa; median as cut-off). D) Gene set enrichment analysis showing pathways of interest (in the top 10 pathways ordered by FDR) enriched between TIL-High PCa vs. TIL-Low PCa. (FDR - False Discovery Rate, NES - Normalised Enrichment Score).

Figure 6. Identification of SNVs and CNAs in FFPE radical prostatectomy samples using a PCa-specific panel

A) Schematic showing the panel design and analysis pipeline. B) Oncoprint summary of somatic mutations and copy number alterations in 18 patients. C) Putative copy number alterations in *PTEN*, *CHD1* and *RB1* were validated using Taqman real-time PCR.

Figure 7. Exome sequencing of selected cases

A) SNVs and short indels in 5 samples sequenced by exome capture. B) Immune cell counts in the same 5 samples, showing different sub-types quantified on the Visiopharm system. C) Number of neoantigens predicted (NetMHCPan/NetMHCIIpan binding affinity < 500 nM) D) Relationship between TIL counts and total number of coding mutations (no significant correlation).

Figure 8. Mutations in mismatch repair genes

A) Missense mutations in *MLH1* and *MSH6* identified by targeted smMIP sequencing. B) H & E immunohistochemistry for MLH1 and MSH6, showing loss of staining in MSH6 but not MLH1 (dashed line denotes tumour region). Staining pattern in a positive control is also shown. c) STR analysis for MSI markers Bat25 and Bat26.

Table 1. A summary of antibodies used for immunohistochemistry and specific epitope retrieval conditions

Primary Antibody	Type	Supplier	Cat #	Antigen Retrieval	Dilution
CD3 (LN10)	M	Leica Biosystems	PA0553	20 min ER2	RTU
CD4 (LB12)	M	Leica Biosystems	PA0427	20 min ER2	RTU
CD8 (4B11)	M	Leica Biosystems	PA0183	20 min ER2	RTU
CD20 (L26)	M	Leica Biosystems	PA0200	20 min ER1	RTU
PDL-1 (E1L3N(R))	M	Cell Signaling Technology	13684S	20 min ER2	1/200
FOXP3 (236A/E7)	M	Abcam	Ab96048	20 min ER1	1/100
Granzyme B (11F1)	M	Leica Biosystems	PA0291	20 min ER2	RTU
Cytokeratin 5 (XM26)	M	Leica Biosystems	PA0468	20 min ER2	RTU
Multicytokeratin (AE1/AE3)	M	Leica Biosystems	PA0909	10 min Enz 1	RTU
Androgen Receptor	P	Abcam	ab74272	20 min ER1	1/200
ERG (EP111)	M	DAKO	M7314	20 min ER1	1/100
PTEN (Y184)	M	Abcam	ab32199	20 min ER1	1/100
Mib-1 (ki-67)	M	DAKO		20 min ER2	1/400

(M – monoclonal, P – polyclonal)

CD3	T cell
CD4	T helper cell
CD8	Cytotoxic T cell
CD20	B cell
PDL-1	PD1 ligand
FOXP3	T regulatory cell
Granzyme B	Cytotoxic T cell

Table 2: List of fusions identified by RNA-seq analysis

Sample	Fusions
EU_0102	*TMPRSS2*:*ERG*
EU_0304	*TMPRSS2*:*ERG*, *SLC45A2*:*AMACR*
EU_0506	*TMPRSS2*:*ERG*
EU_0708	
EU_0910	*TMPRSS2*:*ERG*
EU_1112	
EU_1314	
EU_1516	*TMPRSS2*:*ERG*
EU_A	*TMPRSS2*:*ERG*, *EIF4E3*:*FOXP1*
EU_B	
EU_C	
EU_D	*TMPRSS2*:*ERG*
EU_E	
EU_F	
EU_G	
EU_H	*TMPRSS2*:*ETV4*
EU_I	
EU_J	*SLC45A3*:*ERG*, *PDZRN3*:*FOXP1*
EU_K	*EIF4E3*:*FOXP1*
EU_L	*TMPRSS2*:*ERG*
EU_M	*TMPRSS2*:*ERG*
EU_N	
EU_O	*TMPRSS2*:*ERG*
EU_P	
EU_Q	*EIF4E3*:*FOXP1*
EU_R	
EU_S	

Supplementary figure legends

Supplementary figure 1. Tumour Region of Interest (ROI) detection pipeline

A) The Tissuealign module was used to align 3 digitised serial sections: 2 slides stained with tumour markers CK5 and PanCK another slide stained with a cancer or immune biomarker (e.g. CD3). B) Regions were labelled as Centre of Tumour (CT) – yellow, Invasive Margin (IM) – magenta, benign regions based on CK5 – red, the entire tissue region (to separate it from the background) – dark blue, regions that are excluded from the analysis – cyan. C) An example of ROI labelling in the tumour microarray.

Supplementary figure 2. smMIP panel sequencing QC

The graph shows the percentage of MIPs with the indicated depth of coverage (reads) for each sample analysed. NTC = negative control.

Supplementary figure 3. TMPRSS2:ERG fusion transcripts

Schematic illustrations of the exons involved in TMPRSS2:ERG fusions discovered in this cohort. The vertical lines indicate the exons and the transcripts TMPRSS2 and ERG are depicted on the left and right respectively, with their respective ENSEMBL transcript IDs.

Supplementary figure 4. Association between ERG status and lymphocyte density in the TCGA and CamCap datasets

A) H&E tumour images from the The Cancer Genome Atlas (TCGA) prostate adenocarcinoma (PRAD) samples were analysed for lymphocyte density using our pipeline. TMPRSS2:ERG fusion status was obtained from the Genomic Data Commons portal. B) Tissue microarrays from the CamCap cohort were analysed for ERG status by IHC and lymphocyte density was measured using our pipeline. (ns = not significant, $p > 0.05$, Welch's two sample t test)

Supplementary figure 5. Copy number alterations in 5 samples profiled by whole exome sequencing

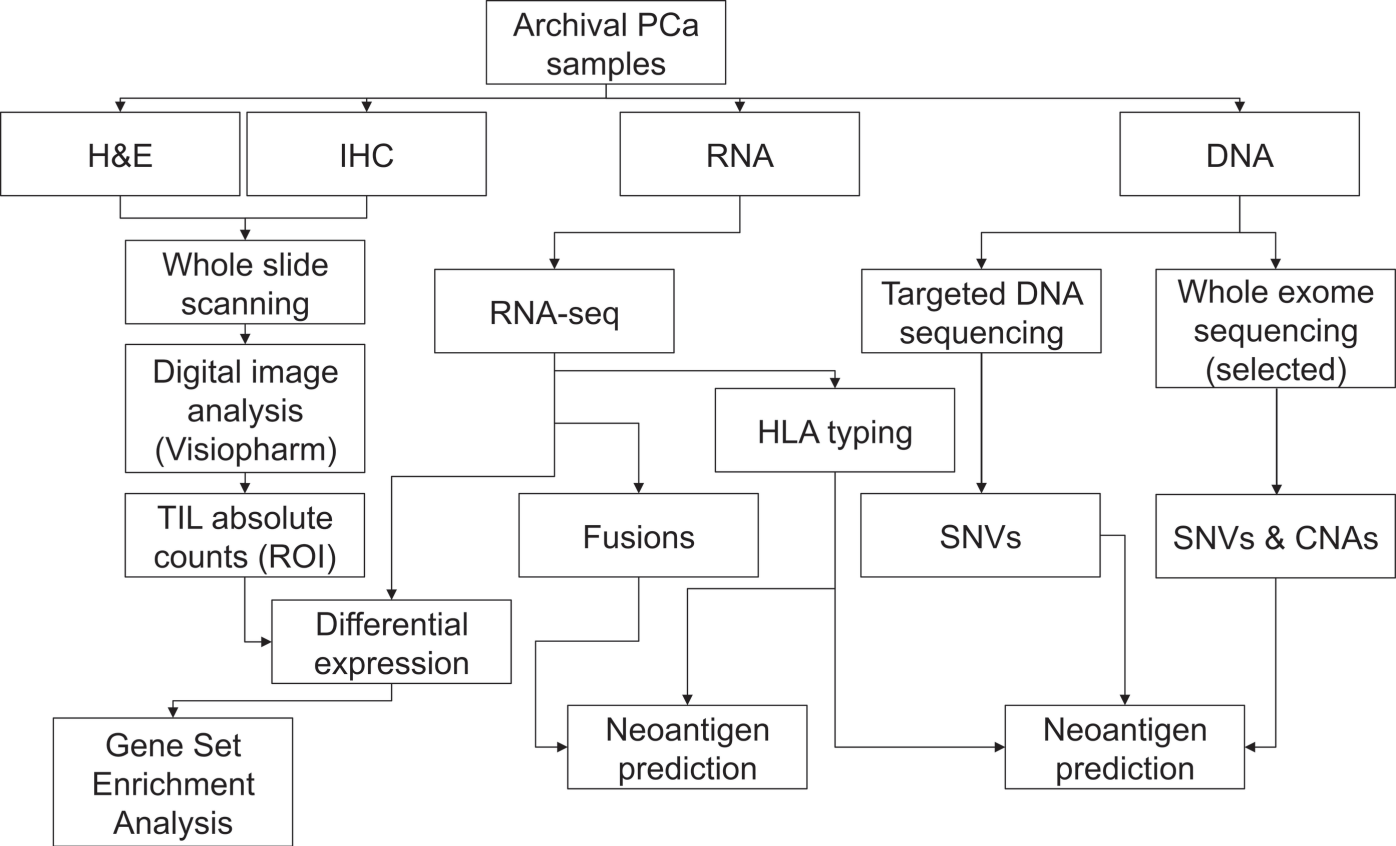
Copy number alterations are depicted as denoised copy ratios (compared to a panel of normals) along with their respective alternate allele frequencies for each sample.

Supplementary figure 6. Immune cell infiltration centred around the benign glands

Benign prostate glands were found to have immune cell infiltration around them, as seen by Eosin and Hematoxylin staining.

Supplementary figure 7. Case EU_O with TMPRSS2:ERG fusion by RNA_seq but negative for ERG on IHC

ERG IHC showing the relatively small tumour glands used for assessment (arrow marks tumour glands).


Archival PCa samples
H&E
IHC
RNA
DNA
Whole slide scanning
Digital image analysis (Visiopharm)
TIL absolute counts (ROI)
RNA-seq
HLA typing
Fusions
Differential expression
Gene Set Enrichment Analysis
Neoantigen prediction
Targeted DNA sequencing
SNVs
Neoantigen prediction
Whole exome sequencing (selected)
SNVs & CNAs

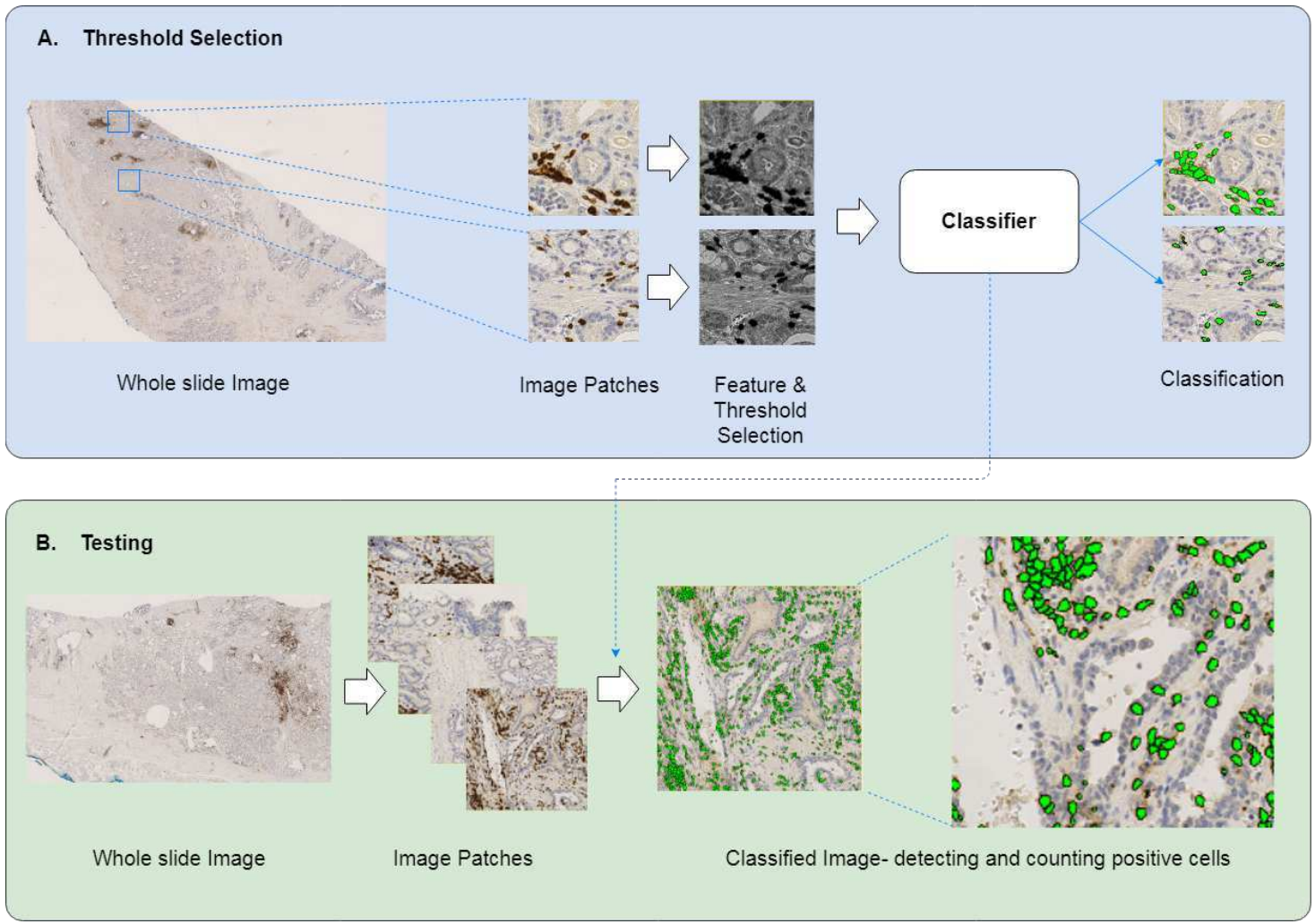


Figure 2. Image analysis pipeline to quantify immune cells.

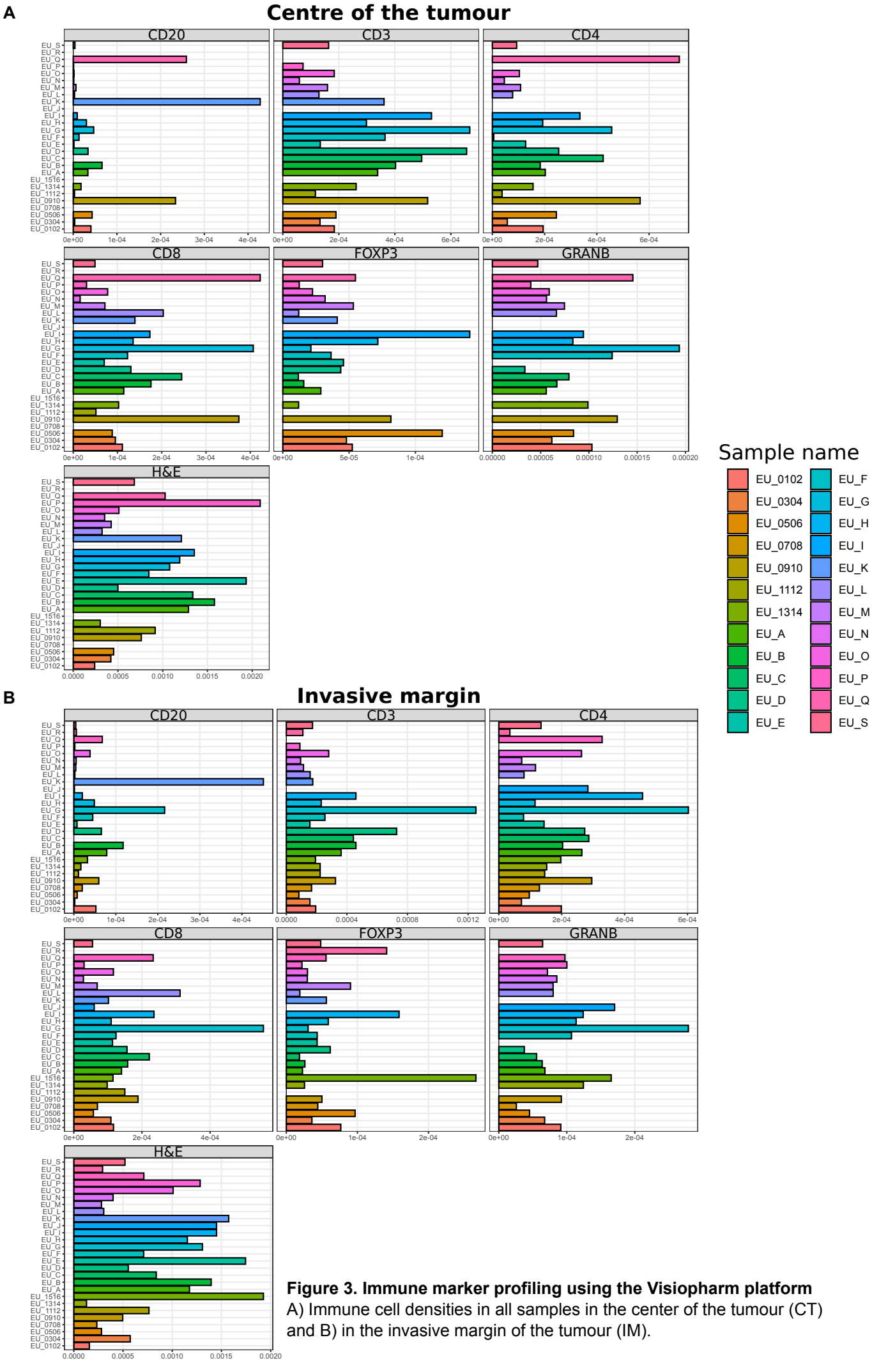


Figure 3. Immune marker profiling using the Visiopharm platform
A) Immune cell densities in all samples in the center of the tumour (CT) and B) in the invasive margin of the tumour (IM).

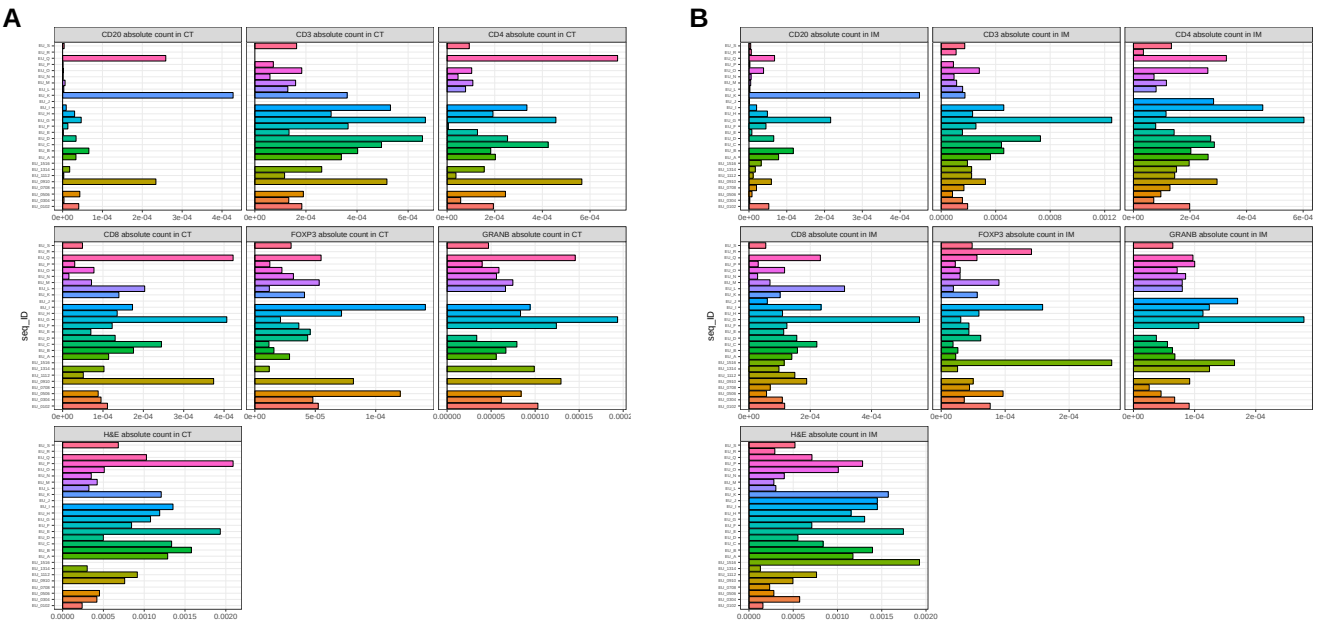


Figure 3. Immune marker profiling using the Visiopharm platform

A) Absolute immune cell counts in all samples in the center of the tumour (CT) and B) in the invasive margin of the tumour (IM).

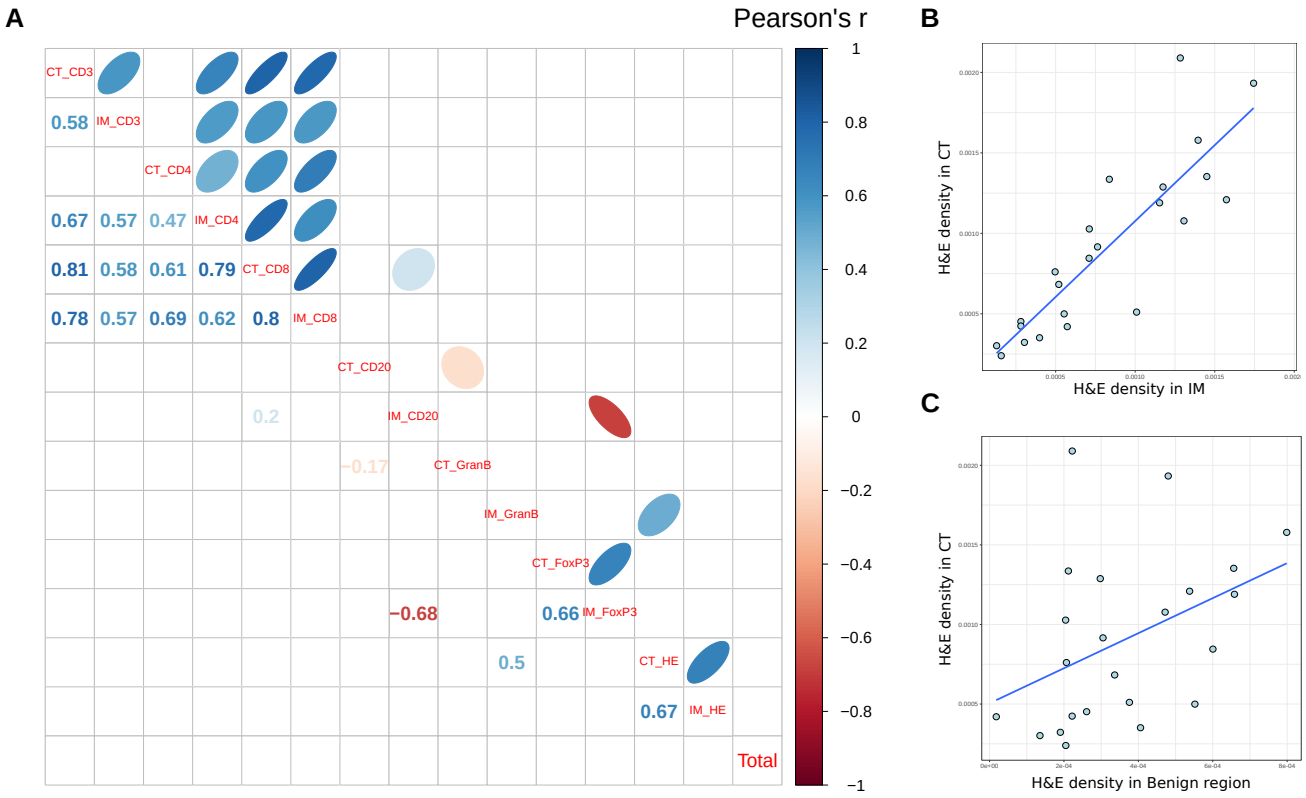


Figure 4. Correlations between immune cell subtypes in different regions of the tumour
A) Correlogram of immune cell subtypes identified. Blue or red colour scale indicates positive or negative correlation respectively, with only those correlations with a p value < 0.05 (not adjusted for multiple comparisons) being shown. B) Relationship between immune cell densities in IM vs. CT (Pearson's R = 0.86, p < 0.001) and C) benign region vs. CT (Pearson's R = 0.41, p = 0.05) shown as scatter plots.

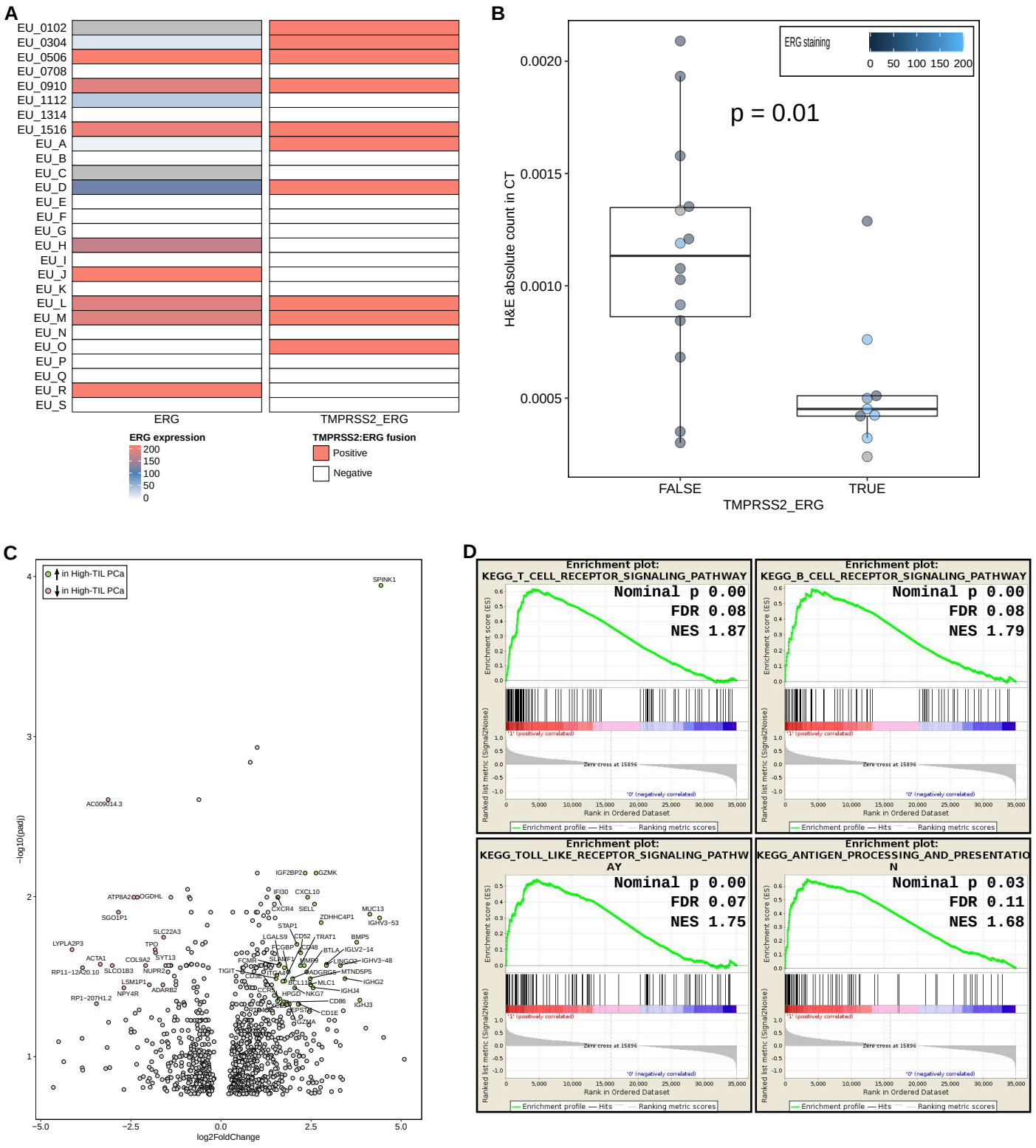


Figure 5. Corroboration of TMPRSS2:ERG fusions with ERG staining, and differences in immune cell counts

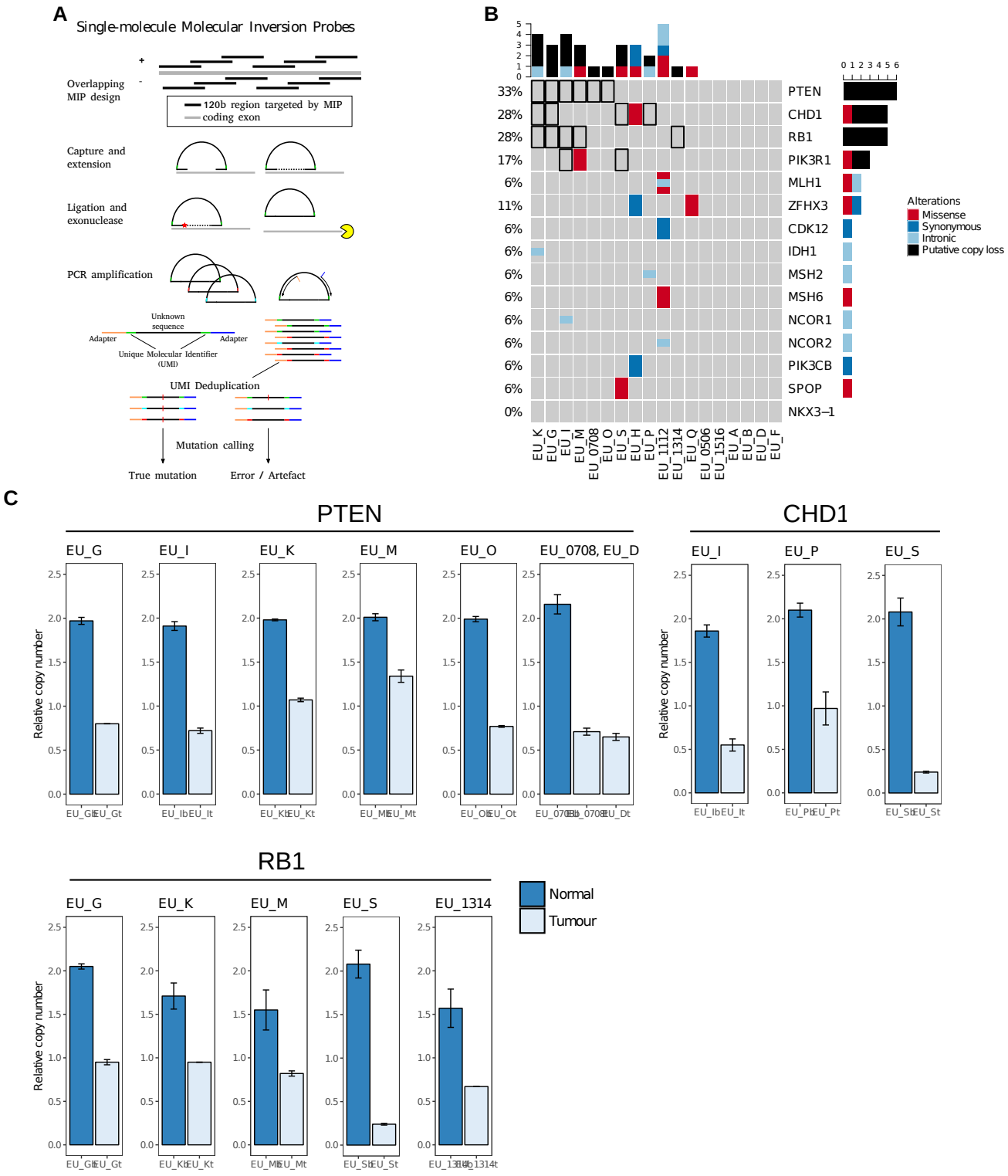


Figure 6. Identification of SNVs and CNAs in FFPE radical prostatectomy samples using a PCa-specific panel

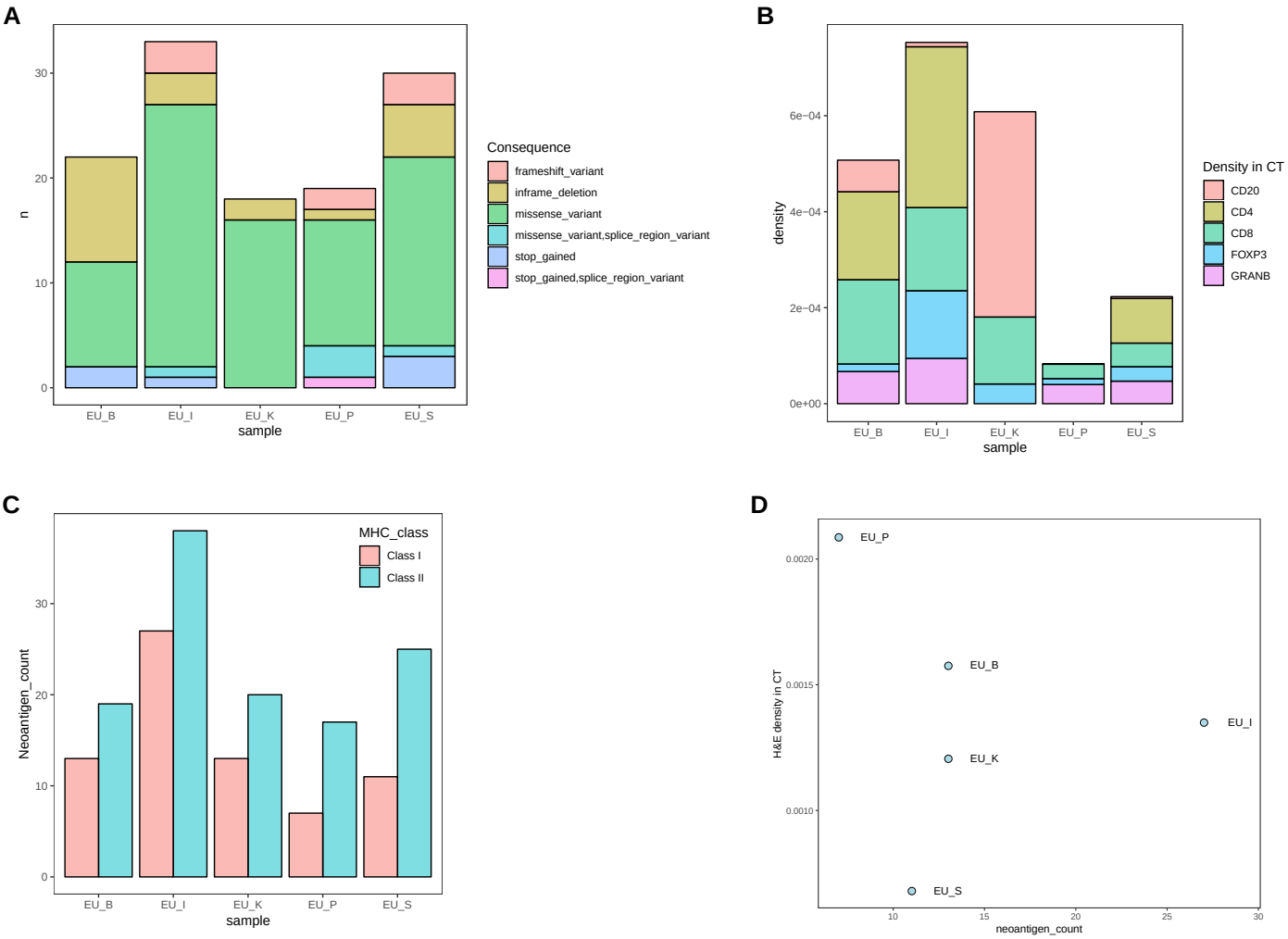
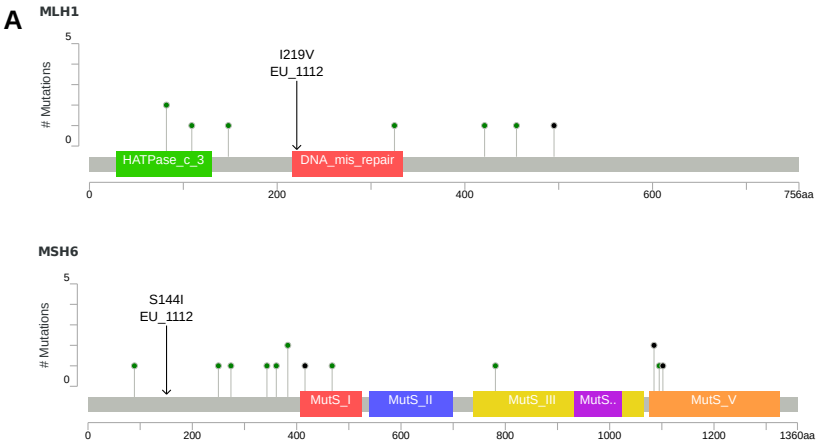
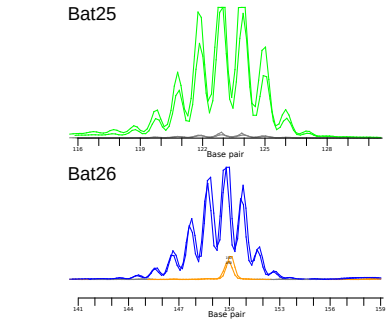


Figure 7. Exome sequencing of selected cases.





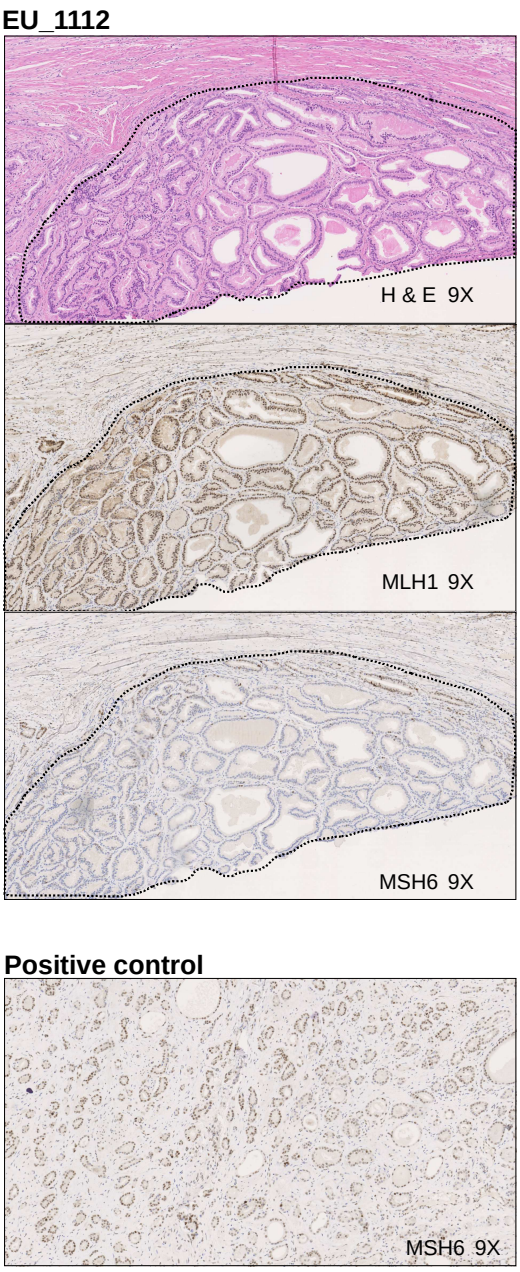
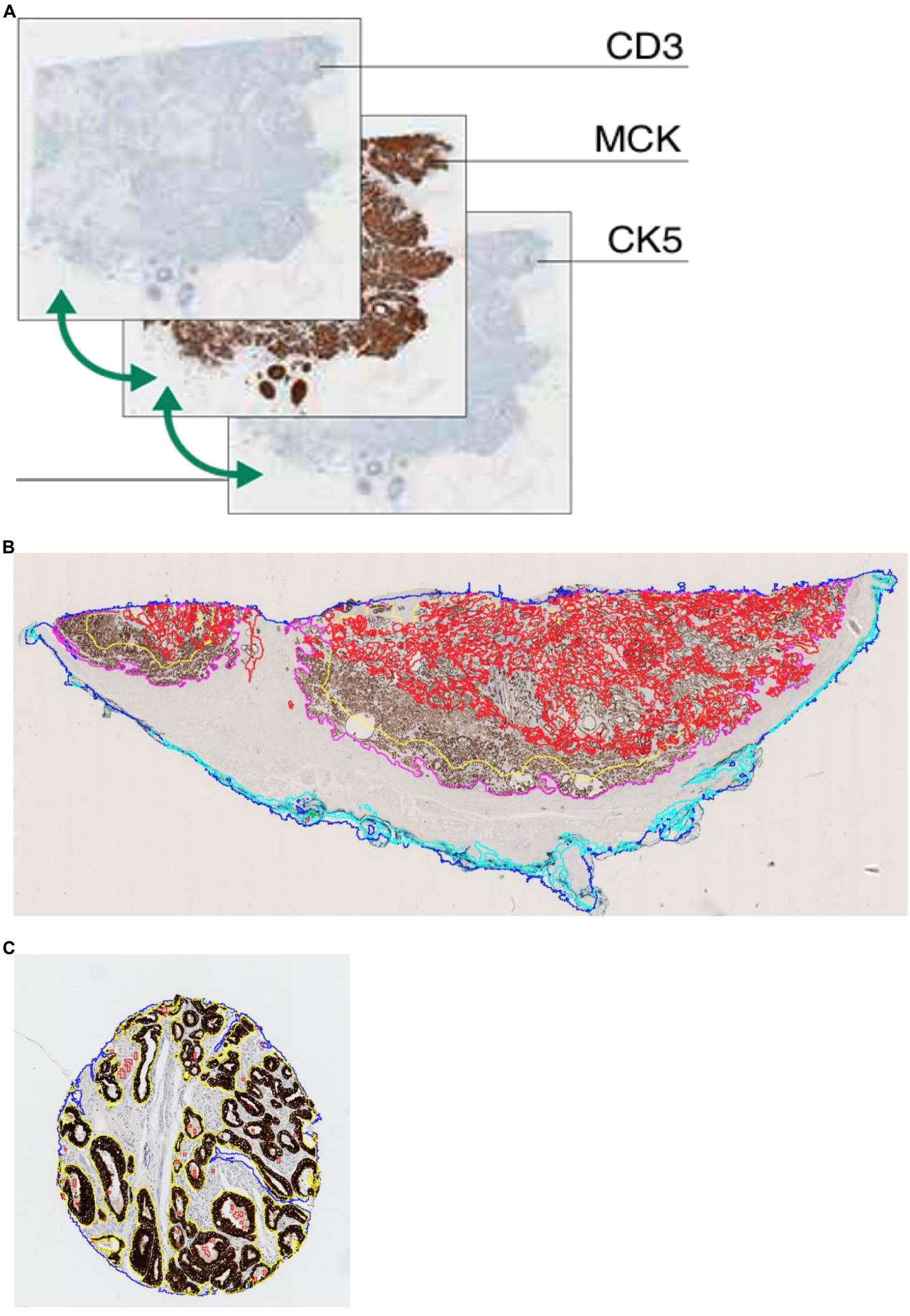
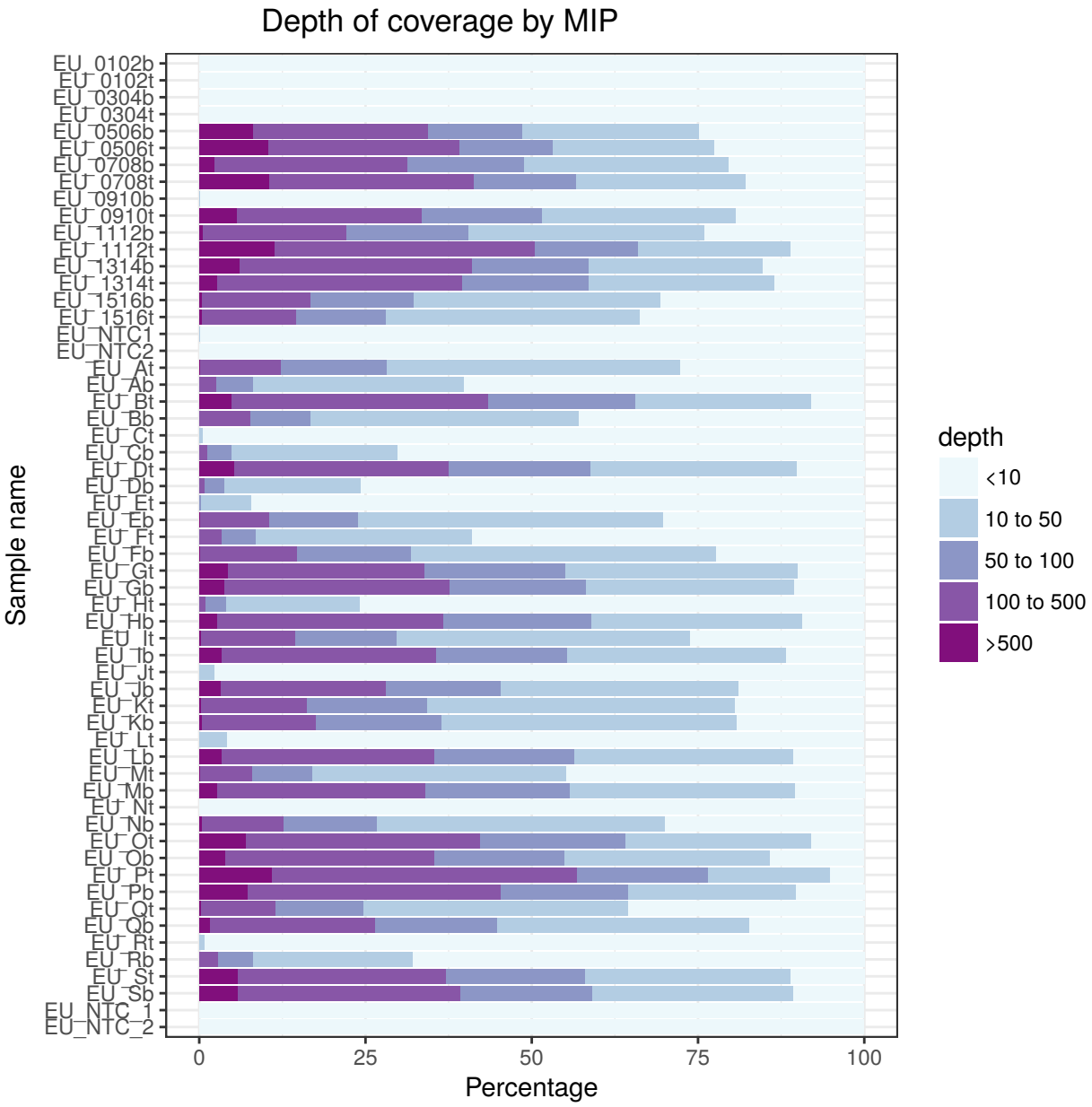


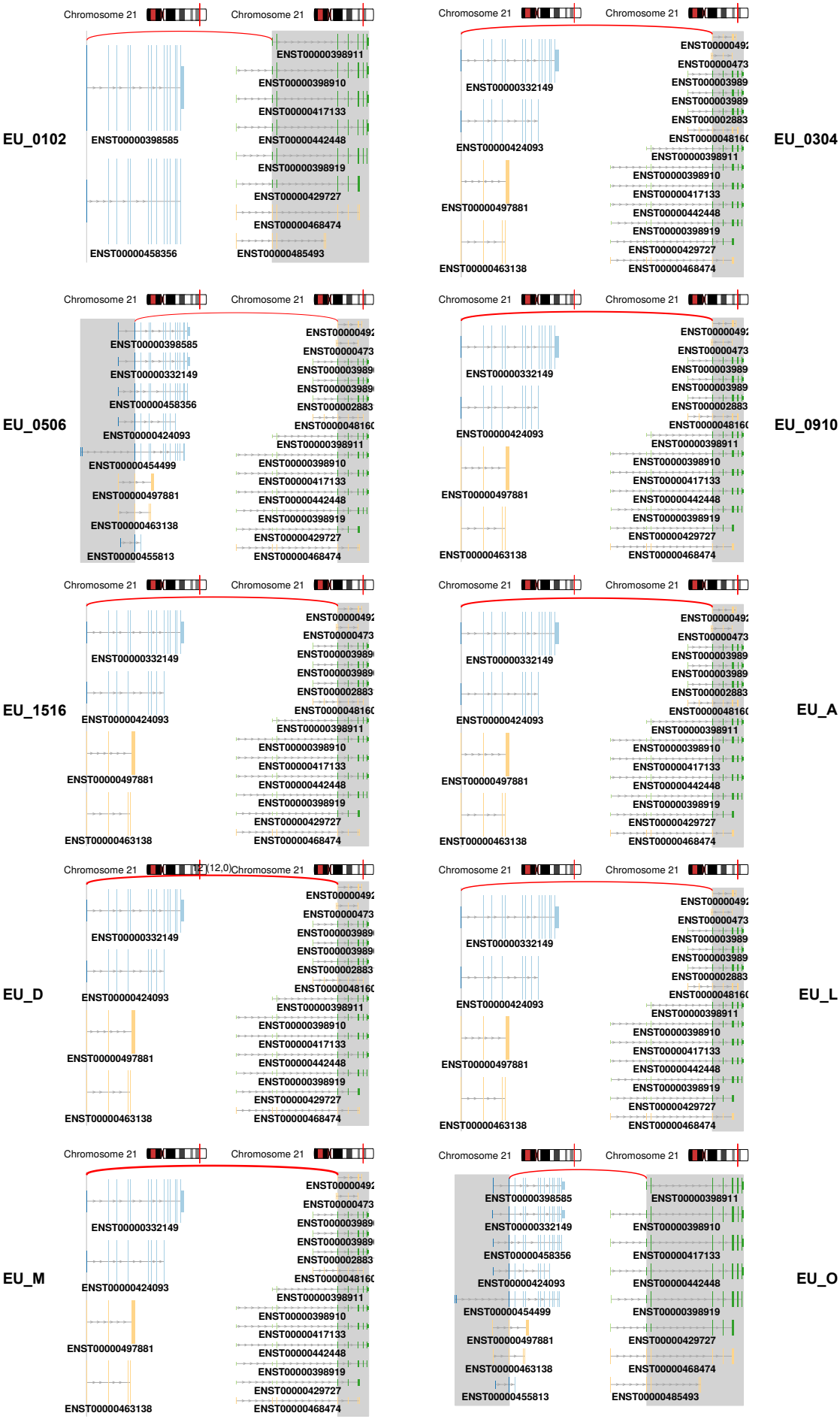
Figure 8. Mutations in mismatch repair genes



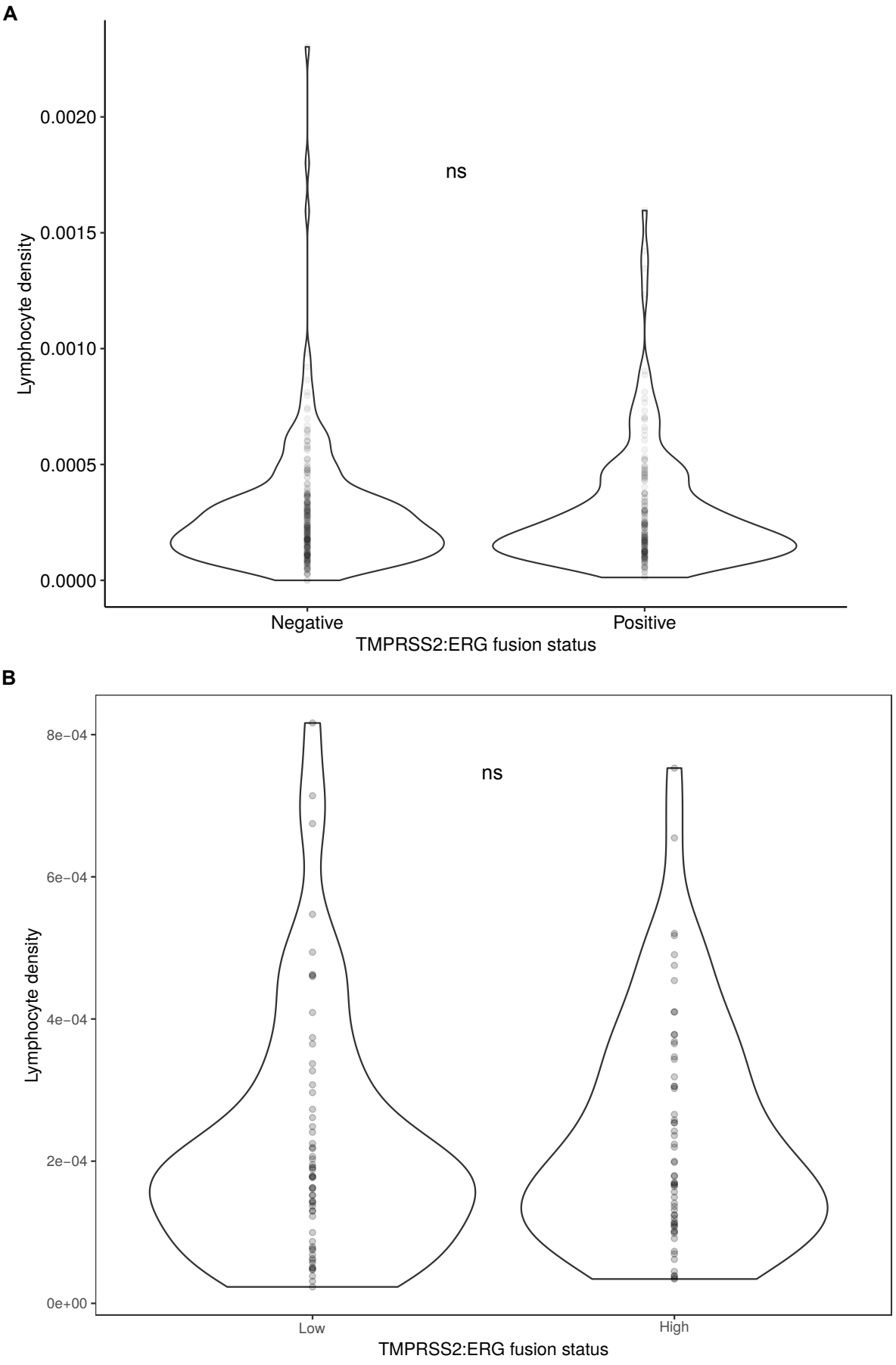
Supplemental figure 1. Tumour ROI pipeline



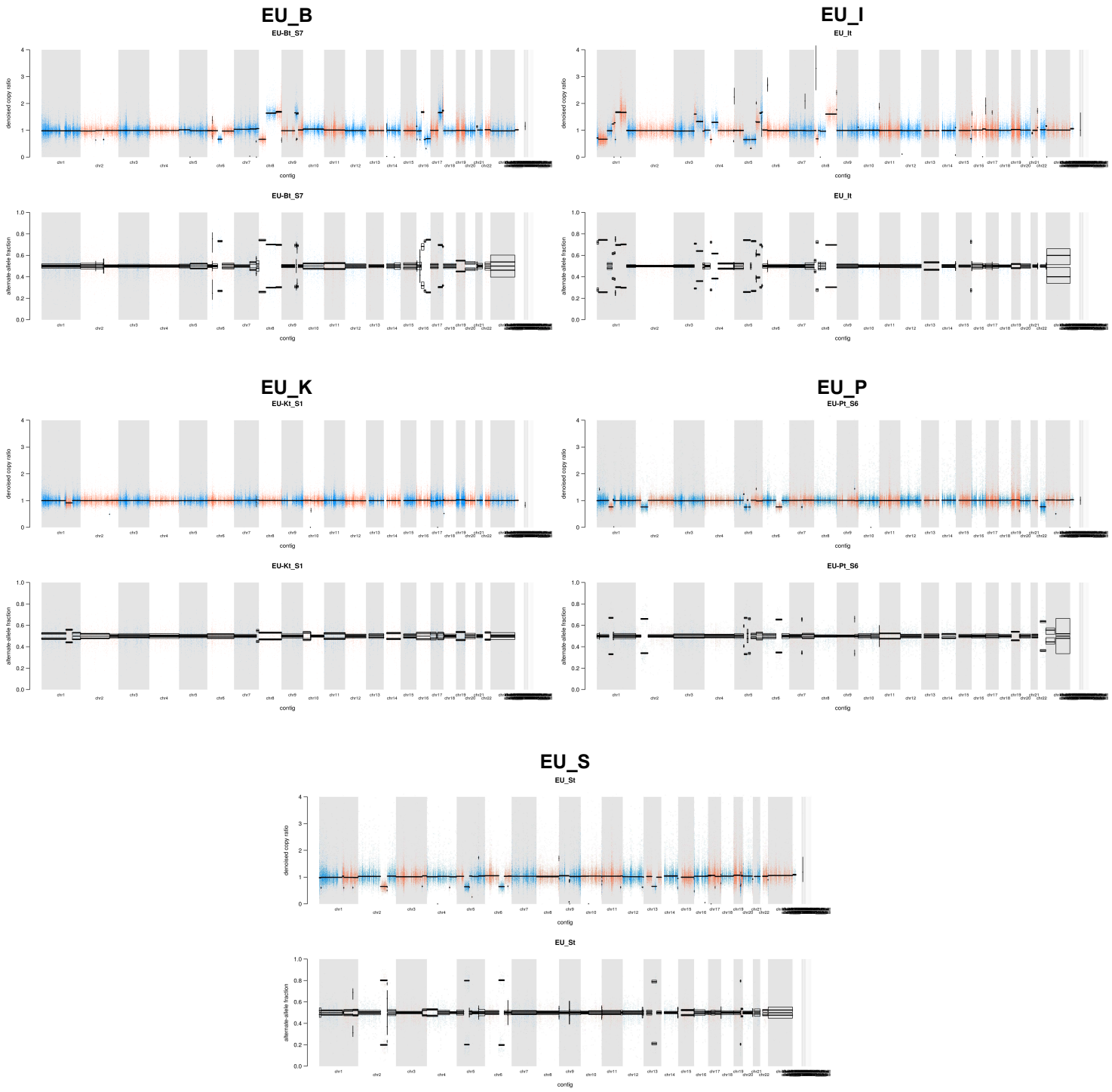
Supplemental figure 2. smMIP panel sequencing QC



Supplemental figure 3. TMPRSS2:ERG fusion transcripts

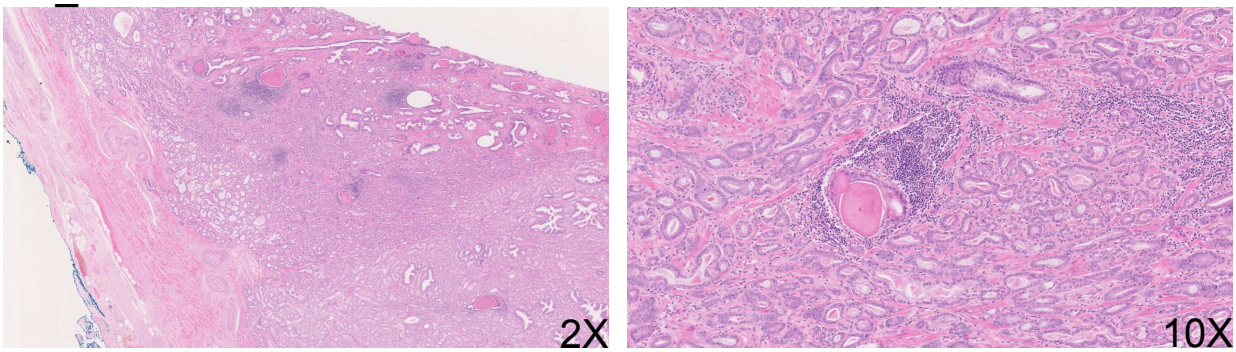


Supplemental figure 7. Association between ERG status and lymphocyte density in the TCGA and CamCap datasets.

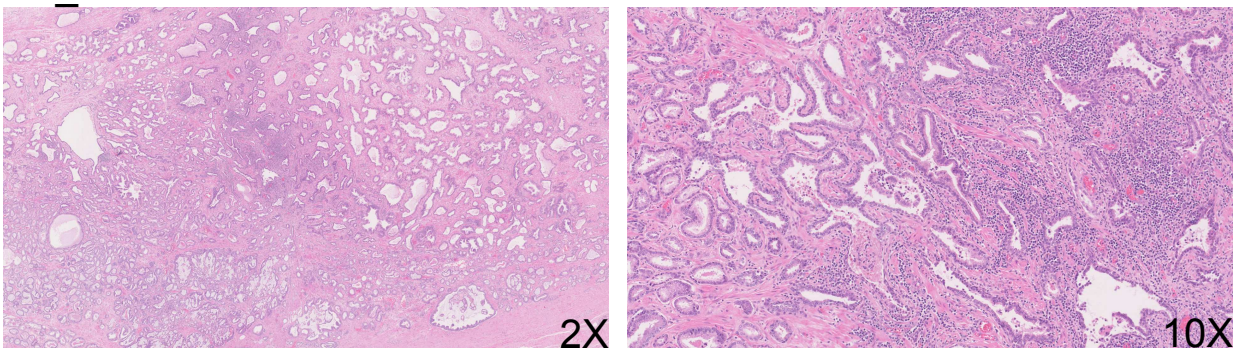


Supplemental figure 4. Copy number alterations in 5 samples profiled by whole exome sequencing

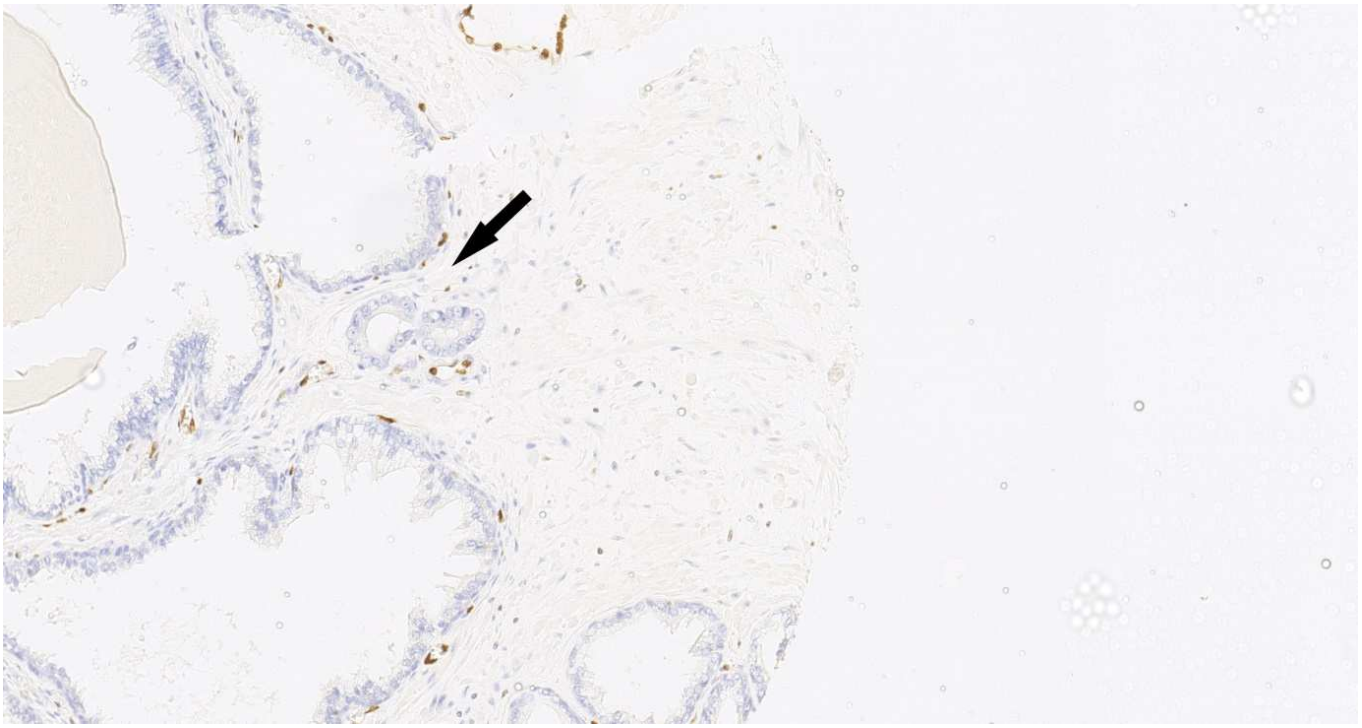
EU_0910



EU_0506



Supplemental figure 5. Immune cell infiltration centred around the benign glands



Supplemental figure 6. Case EU O with TMPRSS2:ERG fusion by RNA Seq but negative for ERG on IHC

Supplementary table 1: Comparison with manual counts

Marker	Manual count	Visiopharm 40x count	Visiopharm 10x count
CD3	121	107	116
CD20	67	70	66
CD4	62	61	63
CD8	49	46	51
FoxP3	33	34	34
GranB	28	30	30
H&E	384	394	390

Supplementary table 2: Morphological assessment

Reference number	Budding Y=1, N=0	Infiltrative =1, Circumscribed = 0	Stage	Gleason
EU_J	0	0	2	3+4
EU_L	0	0	3a	3+4
EU_D	1	0	2	3+4
EU_M	0	0	2	3+4
EU_H	0	0	2	3+4
EU_Q	1	1	2	4+3
EU_B	0	0	3b	4+3 with tertiary pattern 5
EU_I	0	0	3b	3+4
EU_S	0	0	2	4+3
EU_G	0	1	3a	3+4
EU_R	0	0	3a	3+4
EU_A	0	1	3a	4+3
EU_F	0	0	2	3+4
EU_P	1	1	3b	4+5
EU_O	1	1	3a	3+4
EU_N	0	0	3a	4+3
EU_K	0	0	3a	3+4
EU_E	0	0	2	3+4
EU_C	0	0	2	3+4
EU09	0	0	3a	3+4
EU07	0	0	3a	4+5
EU01	0	0	2	3+4
EU05	0	0	3a	4+3
EU03	1	1	3a	3+4
EU15	0	1	3a	4+3
EU13	0	0	2	3+4
EU11	0	0	3a	3+4

Supplementary table 3: IHC of tissue microarray

Reference	Proliferation index Mib-1	ERG Score 0-3	PDL-1 Score 0,1,2,3	AR H Score	PTEN H Score
EU_A	0	1	0	0	200
EU_C	0	0	0	x	x
EU_K	1	0	0	20	300
EU07	1	0	0	30	100
EU_B	0	0	1	0	200
EU_D	0	1	0	20	200
EU_J	5	2	0	300	150
EU01	x	x	0	x	x
EU09	0	2	0	30	0
EU15	5	3	1	100	195
EU05	1	3	0	0	300
EU03	1	1	0	40	5
EUFt	0	0	0	0	100
EU13	8	0	0	80	5
EU_N	0	0	0	0	300
EU_O	5	0	0	50	300
EU_G	0	0	1	0	200
EU_I	5	0	0	0	5
EU_P	5	0	0	0	100
EU_R	5	2	2	150	200
EU_H	0	3	0	80	300
EU_M	5	3	0	40	40
EU_Q	0	0	0	80	90
EU_E	0	0	0	80	200
EU11	1	2	0	0	100
EU_L	5	3	0	90	70
EU_S	2	0	0	0	200

Supplementary table 4: Predicted neoantigens

Neoantigen score for TMPRSS2:ERG fusion peptides for MHC Class I alleles

sample	Gene Name	HLA Allele	Peptide Length	MT Epitope Seq	NetMHCpan MT Score
EU_03	TMPRSS2>>ERG	HLA-B*35:01	10	FPNCPCLLTF	10.5
EU_05	TMPRSS2>>ERG	HLA-B*07:26	10	FPNCPCLLTF	341.7
EU_09	TMPRSS2>>ERG	HLA-B*40:02	9	YETMLTHVL	7.9
EU_M	TMPRSS2>>ERG	HLA-A*02:36	10	LLYETMLTHV	23.6
EU_M	TMPRSS2>>ERG	HLA-B*07:02	10	FPNCPCLLTF	189.6
EU_O	TMPRSS2>>ERG	HLA-B*08:01	10	FPNCIRVHTM	58.8

Neoantigen score for TMPRSS2:ERG fusion peptides for MHC Class II alleles

sample	Gene Name	HLA Allele	Peptide Length	MT Epitope Seq	NetMHCIIpan MT Score
EU_01	TMPRSS2>>ERG	HLA-DRB1*07:01	15	TCRLYKQSVSRLLYE	9.68
EU_09	TMPRSS2>>ERG	HLA-DPA1*01:03/DPB1*60:01	15	SVSRLLYETMLTHVL	13.03
EU_M	TMPRSS2>>ERG	HLA-DRB1*12:01	15	TCRLYKQSVSRLLYE	5.63
EU_M	TMPRSS2>>ERG	HLA-DRB1*12:01	15	NCPCLLTFCCHPSGI	162.57
EU_O	TMPRSS2>>ERG	HLA-DPA1*01:03/DPB1*104:01	15	KYALADFTNRVFPNC	115.3

Supplementary table 5: DE genes in TIL-High vs. TIL-Low

Top 50 (by fold change) differentially expressed genes between TIL-High (positive log2FoldChange) and TIL-Low (negative log2FoldChange) PCa samples. (padj = adjusted p value)

	gene_symbol	baseMean	log2FoldChange	pvalue	padj
1	SOX9	4.222591	19.91171	1.51E-11	3.21E-07
2	SPINK1	189.9429	4.430591	1.07E-08	0.000113
3	IGHV3-53	5.796331	4.39594	2.30E-05	0.013506
4	LYPLA2P3	3.237906	-4.13859	5.33E-05	0.021306
5	MUC13	80.90676	4.1234	1.94E-05	0.012803
6	IGHJ3	6.071059	3.846454	0.000324	0.044033
7	RP11-12A20.10	2.153347	-3.8463	0.00013	0.027656
8	BMP5	45.73508	3.759554	3.89E-05	0.01918
9	RP1-207H1.2	5.505191	-3.45897	0.000371	0.046371
10	IGHG2	454.3003	3.431158	0.000182	0.032345
11	ACTA1	38.08311	-3.35421	8.78E-05	0.026304
12	IGHV3-48	9.479455	3.307607	0.000117	0.026767
13	AC009014.3	8.464434	-3.13552	6.98E-07	0.002464
14	SLCO1B3	47.47918	-3.02279	0.000107	0.026767
15	IGLV2-14	50.63769	2.922311	8.95E-05	0.026304
16	LINGO2	3.930789	2.918528	9.80E-05	0.026767
17	SGO1P1	3.994004	-2.85009	1.76E-05	0.012435
18	ZDHHC4P1	6.448396	2.773569	2.59E-05	0.014444
19	NPY4R	28.92077	-2.69979	0.00025	0.03691
20	GZMK	18.342	2.633104	2.44E-06	0.007083
21	SELL	68.78225	2.589032	1.17E-05	0.011066
22	IGHJ4	17.78192	2.551094	0.000255	0.03691
23	MLC1	8.210756	2.482832	0.000221	0.035361
24	MTND5P5	5.194503	2.473293	0.000183	0.032345
25	ATP8A2	90.55514	-2.42375	8.98E-06	0.01003
26	CXCL10	70.17327	2.397759	9.00E-06	0.01003
27	BTLA	27.49205	2.372027	0.000155	0.029362
28	IGF2BP2	110.1526	2.330304	2.87E-06	0.007083
29	OGDHL	36.4636	-2.32526	5.86E-06	0.01003
30	MMP9	24.24749	2.304347	9.89E-05	0.026767
31	CD52	9.614972	2.20557	6.16E-05	0.022286
32	TRAT1	20.50831	2.199969	0.000102	0.026767
33	CD1E	7.634753	2.153609	0.000403	0.046747
34	EPSTI1	76.91579	2.135403	0.000387	0.046747
35	STAP1	10.86657	2.100702	4.29E-05	0.019726
36	COL9A2	259.4917	-2.0968	0.000118	0.026767
37	NKG7	5.958249	2.040286	0.000251	0.03691
38	LSM1P1	3.043951	-1.99365	0.000225	0.035361
39	ADGRG5	12.17984	1.981169	0.000178	0.032345
40	GZMA	6.676635	1.889664	0.0004	0.046747
41	FCGBP	182.8459	1.861101	0.000155	0.029362
42	TPO	223.6325	-1.83187	5.10E-05	0.021306
43	SYT13	67.94423	-1.82028	6.41E-05	0.022286
44	CXCL9	52.10411	1.819257	0.000409	0.047106
45	CD86	27.88841	1.81805	0.000359	0.045517
46	CD48	39.77622	1.761854	0.000191	0.033434
47	SLAMF1	20.82475	1.754308	0.000137	0.027656
48	HPGD	304.4284	1.73572	0.000343	0.045154
49	BCL11B	66.41537	1.708898	0.000203	0.033604
50	RTP4	10.34162	1.687188	0.000404	0.046747

Supplementary table 6: DE genes in TMPRSS2:ERG fusion-positive vs. fusion-negative

Top 50 (by fold change) differentially expressed genes between TMPRSS2:ERG fusion-positive (positive log2FoldChange) and fusion-negative (negative log2FoldChange) PCa samples. (padj = adjusted p value)

	gene_symbol	baseMean	log2FoldChange	pvalue	padj
1	OR4K12P	31.07744	-8.1252	1.50E-09	8.43E-06
2	BX072566.1	29.44694	-6.07578	0.000394	0.042153
3	RP11-162O14.2	7.201725	-5.71551	0.000397	0.042221
4	GLDC	4.200217	4.529977	3.40E-06	0.001855
5	CH17-212P11.4	2.319084	-4.3754	0.000319	0.037708
6	POTEB3	30.49017	-4.37529	1.83E-05	0.005736
7	RP11-403B2.10	5.042045	-4.30739	0.00013	0.021577
8	NF1P2	13.72844	-4.29663	4.73E-06	0.002218
9	RPS3AP54	3.890426	4.129026	1.53E-10	1.37E-06
10	ATOH1	1.938748	-4.11418	0.00026	0.033349
11	RP11-983G14.1	7.38755	-4.04284	1.62E-05	0.005227
12	SLC9C2	4.223327	-3.99862	0.000142	0.023105
13	RP11-810K23.10	18.3833	-3.91954	7.62E-06	0.002963
14	RP11-390F4.2	1.861183	3.90548	6.89E-09	2.38E-05
15	RLN1	70.05678	-3.86248	2.83E-05	0.007508
16	COL2A1	120.8565	3.840519	1.59E-05	0.005202
17	MED15P7	8.439892	-3.63639	3.75E-05	0.009187
18	SPINK1	96.79463	-3.60421	1.69E-06	0.001159
19	RP11-810K23.9	28.05059	-3.37277	9.82E-05	0.018152
20	BMP5	48.90497	-3.34981	0.000249	0.032649
21	TFF3	137.6855	-3.33697	4.79E-06	0.002218
22	NF1P9	6.038832	-3.32607	0.000512	0.049319
23	ANKRD34B	36.41605	3.314607	0.000343	0.039405
24	RP1-207H1.2	4.918794	3.287518	4.51E-05	0.010276
25	ERG	937.1985	3.233903	1.24E-08	3.07E-05
26	AC009014.3	9.357759	3.22067	1.82E-10	1.37E-06
27	RP11-191L9.5	5.660824	3.173773	2.90E-06	0.001769
28	MUC13	21.02118	-3.01212	7.54E-05	0.014788
29	CHRM3	324.8776	2.821299	1.50E-08	3.07E-05
30	TERT	2.373596	2.814404	6.51E-05	0.013224
31	FTH1P20	20.53278	2.778275	1.24E-05	0.004358
32	RP11-132A1.6	7.133729	-2.59803	1.57E-05	0.005194
33	TRPC7	4.071373	2.526416	5.16E-05	0.011185
34	TDRD1	314.5464	2.503937	3.66E-07	0.000454
35	AC013410.1	2.173015	2.495773	5.78E-05	0.012074
36	SERPINA11	5.481057	2.49203	3.64E-05	0.00912
37	MDFI	43.64216	2.468747	9.51E-08	0.000143
38	RNF2P1	8.323269	2.442392	3.96E-05	0.009558
39	SLC29A3	5.871698	-2.4282	0.000424	0.044524
40	DACT2	65.16056	2.419653	3.62E-06	0.001891
41	GAPDHP71	4.147833	2.337177	0.0003	0.036221
42	NPY4R	27.89688	2.318063	0.000236	0.031826
43	CTB-167G5.7	8.815664	2.310765	0.000383	0.041562
44	OGDHL	38.29	2.306558	2.06E-07	0.000291
45	GMNC	9.989964	-2.29828	0.000344	0.039405
46	C22orf42	8.976664	2.294712	0.000171	0.025663
47	GZMK	16.74993	-2.24074	0.000198	0.028888
48	ACTN2	23.99649	-2.20777	0.000434	0.044524
49	SYT13	70.5699	2.190136	1.24E-08	3.07E-05
50	C15orf48	38.98175	-2.17798	0.000214	0.030415

Supplementary table 7: Copy number alterations in WES

Copy number alterations in key prostate cancer driver genes, called from whole exome sequencing data of 5 PCa samples

	Copy loss	Copy gain
EU_B	CDH1	MYC
EU_I	CHD1, PIK3R1	MYC
EU_K		
EU_P	CHD1, PIK3R1	
EU_S	CHD1, PIK3R1	MAP3K1